M

Macfie, Alec Lawrence (1898–1980)

Andrew Skinner

Alec Macfie was born in Partick on 29 May 1898. He went first to school at Hillhead but later joined his brother at the High School of Glasgow where he had a distinguished career.

When he left school, Macfie, too young to enlist, worked in a munitions factory for a few months. But in 1916 he joined the Second Battalion, the Gordon Highlanders, and was commissioned as a lieutenant. He saw action at Passchendaele, and was badly wounded during an action on the Asiego Plateau in the early summer of 1918. After recovering to some extent from his wounds, Macfie entered Glasgow University where he graduated with first class honours in philosophy and English literature in 1922. Thereafter he entered a law office and took his LL.B. But having opted for an academic career, he returned to the university and took a first in economics and politics in 1927.

In the years that followed Macfie held temporary teaching posts in Nottingham and St Andrews (where he deputized for the professor of moral philosophy) before returning to Glasgow in 1932 as lecturer in the Department of Political Economy under W.R. Scott. Scott (the 'chief') died in 1940 and Macfie was invited to take the Adam Smith Chair in 1945.

Side by side with his teaching, Macfie produced three books in the period up to 1945, all of which reflect his interest in philosophy and psychology as well as in economics: *Theories of the Trade Cycle* (1934), *An Essay on Economy and Value* (1936) and *Economic Efficiency and Social Welfare* (1943). It was not until the mid-1950s, only a few years before retiring, that he embarked on a serious study of Adam Smith with special reference to the *Theory of Moral Sentiments*. Following the acquisition of the manuscripts, discovered by J.M. Lothian in 1958, Macfie became one of the driving forces behind the Glasgow edition of the Works and Correspondence of Adam Smith, and acted as co-editor (with D.D. Raphael) of the *Theory of Moral Sentiments* (1976). He also produced a little book which has exerted an enormous influence in this field, *The Individual in Society: Papers on Adam Smith* (1967).

Few modern scholars have been better equipped for the study of Smith. Macfie was a qualified lawyer, with degrees in philosophy, literature, and economics, while Smith was writing at a time when it was possible to think in terms of a global system of thought which might embrace these separate disciplines.

Selected Works

1934. *Theories of the trade cycle*. London: Macmillan. Reprinted, New York: Augustus M. Kelley, 1971.

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

1936. *An essay on economy and value: Being an enquiry into the real nature of economy.* London: Macmillan.
1940. (ed.) *Studies relating to Adam Smith during the last fifty years,* by W.R. Scott. Proceedings of the British Academy.
1943. *Economic efficiency and social welfare.* Oxford: Oxford University Press.
1967. *The individual in society: Papers on Adam Smith.* London: Allen & Unwin.
1976. (ed., with D.D. Raphael). *The theory of moral sentiments*, by Adam Smith. Oxford: Clarendon Press; Vol. I of the Glasgow edn of the Works and Correspondence of Adam Smith.

Machinery Question

Salim Rashid

Keywords

Circulating and fixed capital; Fixed capital; Machinery question; Mill, J.S.; Ricardo, D.; Say's Law; Stewart, D.; Tozer, J.E.; Tucker, J.; Wages fund; Wicksell, J.G.K

JEL Classifications

O

That machinery is of benefit to the manufacturer who introduces it has never been a point of discussion in the history of economics and the machinery question is solely a dispute over whether society benefits from the introduction of machinery, the most pressing social issue being the displacement of labour by machinery and the consequent threat of widespread unemployment. In general terms, the social benefits of machinery were well appreciated by the middle of the 18th century. However, the greatly increased use of machinery at the end of the 18th century gave a new intensity to the debate at the beginning of the 19th century. The analytical tools used by classical economists to tackle this general equilibrium problem were however quite inadequate and it is doubtful whether a deeper understanding of the issue was achieved by the heroic abstractions of the 19th century.

The earliest explicit discussions of machinery appear to be in the pamphlets of John Cary (1695), *A Discourse on Trade.* It was a time when the competitiveness of English industry was being much discussed and John Cary pointed out that England retained her business advantage because of the ability of English manufacturers to invent.

> Tobacco is cut by Engines: Books are printed; Deal Boards are sawn with Mills; Lead is smelted by Wind-Furnaces; all which save the Labour of many Hands, so the Wages of those employed need not be fallen. . . .New Projections are every Day set on Foot to render the making our Woollen Manufactures easy, which should be rendered cheaper by the Contrivance of the Manufacturers, not by falling the Price of Labour: Cheapness creates Expence, and gives fresh Employments, whereby the Poor will be still kept at Work. (Cary 1695, pp. 99–100)

A few years later, in his *Considerations of the East-India Trade* (1701), Henry Martin advocated the import of cheaper cloth from the East Indies by comparing it with the effects of machinery:

> Arts, and Mills, and Engines, which save the labour of Hands, are ways of doing things with less labour, and consequently with labour of less price, tho' the Wages of Men imploy'd to do them shou'd not be abated. The *East-India* Trade procures things with less and cheaper labour than would be necessary to make the like in *England*; it is therefore very likely to be the cause of the invention of Arts, and Mills, and Engines, to save the labour of Hands in other Manufactures. Such things are successively invented to do a great deal of work with little labour of Hands; they are the effects of Necessity and Emulation; every Man must be still inventing himself, or be still advancing to farther perfection upon the invention of other Men . . . (Martin 1701, pp. 589–90)

At this stage the effect of machinery in preserving competitiveness receives primary emphasis. There is as yet no link drawn between high wages and the incentive to create machinery. In the years that followed only the prolific Daniel Defoe paid serious attention to the role of

machinery without making any substantive analytical contribution. Indeed, Defoe even wondered whether machinery were not sometimes an evil because it displaced labour. In parliamentary debates in 1738 on the making of buttons by loom instead of by hand, Henry Archer implicitly subscribed to the full employment and sustainability thesis in a speech of considerable eloquence:

> As to the honourable gentleman's other arguments, drawn from the number of hands employed in the needle-work manufacture . . . it is, in my humble opinion, a very good argument for dismissing this Bill; because, as the manufacture may be carried on by a much fewer number of hands, with equal advantage to our trade in general, those who are employed in the needle-work way, are so many hands taken from other arts and other manufactures, in which they might be employed to much better purpose.

Archer goes on to use an example that was repeated often by classical economists:

> There was a time, Sir, when all the learning of this kingdom, and the rest of Europe, was contained in manuscripts, the writing of which employed great numbers of hands, and took up a vast deal of time in re-copying. But, Sir, how ridiculous would it have been, if on the discovery of the art of printing, the transcribers and copyers of those manuscripts had joined in a petition to the legislature, that it would be pleased to prohibit the art of printing, for the same reason which the honourable gentleman now uses, because great numbers would thereby be deprived of bread! (Archer 1742)

The next advance was stimulated by Montesquieu's claim in *The Spirit of the Laws* that the introduction of machinery was not necessarily beneficial. This provoked Josiah Tucker to provide one of the best statements on the effects of machinery:

> What is the Consequence of this Abridgment of Labour, both regarding the Price of the Goods, and the Number of Persons employed? The Answer is very short and full, *viz*. That the Price of Goods is thereby prodigiously lowered from what otherwise it must have been; and that a much greater Number of Hands are employed. . . .
>
> And the first Step is, that Cheapness, *ceteris paribus* is an Inducement to buy, – and that many Buyers cause a great Demand, – and that a great Demand brings on a great Consumption; – which great Consumption must necessarily employ a vast Variety of Hands, whether the original Material isconsidered, or the Number and Repair of Machines, or the Materials out of which those Machines are made, or the Persons necessarily employed in tending upon and conducting them: Not to mention those Branches of the Manufacture, Package, Porterage, Stationary Articles, and Book-keeping, &c. &c. which must inevitably be performed by human Labour. . . .
>
> That System of Machines, which so greatly reduces the Price of Labour, as to enable the Generality of a People to become Purchasers of the Goods, will in the End, though not immediately, employ more Hands in the Manufacture, than could possibly have found Employment, had no such machines been invented. And every manufacturing Place, when duly considered, is an Evidence in this Point. (Tucker 1757, pp. 241–2)

The subject received little further impetus in the half-century that followed. The tangential discussion of machinery by Adam Smith in the *Wealth of Nations* perhaps contributed to this state of affairs. The only notable treatment is in the lectures of Dugald Stewart at Edinburgh (1858–78), which were very influential as part of an oral tradition, but which were not available in print till the 1860s. Stewart's contribution lay in seeing the machinery question as part of a much larger and more fundamental policy issue – the trade-off between individual losses and social gains. He therefore links together three issues that had hitherto been separately discussed – the creation of large farms, the benefits of enclosures and the use of machinery. In each case Stewart grants that the hardships imposed on individuals were undeniable. He then continues;

> In judging of the policy of such innovations. . . it is absolutely necessary to abstract from the individual hardships that may fall under our notice, and to fix our attention on those general principles which influence the national prosperity. (Stewart 1856, vol. 8, p. 131)

In deciding upon the benefits of introducing machinery, Stewart observes that the material improvement of mankind and the use of machinery are practically inseparable. The policy recommendation was thus unequivocal.

> It is hardly possible to introduce suddenly the smallest innovation into the Political Economy of a State, let it be ever so reasonable, nay, ever so profitable, without incurring some inconveniences. But temporary inconveniences furnish no objection

> to solid improvements. Those which may arise from the sudden introduction of a machine cannot possibly be of long continuance. The workmen will, in all probability, be soon able to turn their industry into some other channel; and they are certainly entitled to every assistance the public can give them, when they are thus forced to change their professional habits. (1856, vol. 8, p. 193)

The severe post-Napoleonic depression contributed greatly to a reconsideration of the effects of machinery and John Barton should perhaps be given most credit for the new interest with his pamphlet, *Observations on the Circumstances which Influence the Condition of the Labouring Classes of Society* (1817). Commenting on the distinction, inherited from Adam Smith, between circulating and fixed capital, Barton pointed out that only the former serves to employ labour – the latter is embodied in machinery. Since it appeared empirically undeniable that progress involved a greater proportionate use of fixed capital Barton argued that the funds for employing labour, or circulating capital, must be subject to proportionate decrease and lead to greater unemployment. Barton was very clear about the role of high wages in inducing the adoption of machinery.

> It is the proportion which the wages of labour at any particular time bear to the whole produce of that labour, which appears to me to determine the appropriation of capital in one way or the other. For if at any time the rate of wages should decline, while the price of goods remained the same, or if goods should rise, while wages remained the same, the profit of the employer would increase, and he would be induced to hire more hands. If, on the other hand, wages should rise, in proportion to commodities, the labour's share in the produce of his own industry would be increased at the expense of his master, who would of course keep as few hands at possible. – He would aim at performing every thing by machinery, rather than by manual labour. (Barton 1817, pp. 17–18)

How far David Ricardo was directly influenced by Barton in reversing his initial optimistic position on the benefits of machinery is unclear, but in the third edition of his *Principles of Political Economy and Taxation* (1821, pp. 388–95) Ricardo tried to justify some of the pessimistic attitudes to machinery by means of a numerical example. To begin with, we have a farmer whose yearly activities can be summarized as follows:

Fixed Capital	7,000
Wages (Circulating)	13,000
	20,000
Profits (10 per cent)	2,000 (used for consumption)
Total	22,000

The circulating capital is said to 'replace the value of 15,000', that is, to provide the required profit of 2,000. In year 1, the capitalist sets half his workers to construct machines. As surplus value arises from circulating capital, the profits of 2,000 arises in equal parts from the workers in farming and the workers in machines:

Fixed Capital (Old)	7,000
Wages (Farming)	6,500
Profits (Farming)	1,000
Wages (Machines)	6,500 (Embodied in machines)
Profits (Machines)	1,000
Total	22,000

If the farmer still spends 2,000 for his own consumption, he is left with 5,500 to spend on wages the next year. In other words, the wage bill falls from 13,000 to 5,500 because of the construction of machines. The gross produce consists of profits, rent and wages, while the net produce consists of profits and rent only. In our case, there is no rent, so the gross produce falls from 15,000 to 7,500 while the net produce stays at 2,000. Ricardo concludes as follows:

> In this case, then, although the net produce will not be diminished in value, although its power of purchasing commodities may be greatly increased, the gross produce will have fallen from a value of 15,000 *l* to a value of 7,500 *l*, and as the power of supporting a population, and employing labour, depends always on the gross produce of a nation, and not on its net produce, there will necessarily be a diminution in the demand for labour, population will become redundant, and the situation of the labouring classes will be that of distress and poverty.

Subsequently, Ricardo concedes that more workers can be employed in producing goods that the capitalist may wish to consume with his increased real power of consumption, but this may not be strong enough to compensate for the initial loss of employment.

> All I wish to prove, is, that the discovery and use of machinery may be attended with a diminution of gross produce; and whenever that is the case, it will be injurious to the labouring class, as some of their number will be thrown out of employment, and population will become redundant, compared with the funds which are to employ it.

There are a number of curious features about Ricardo's analysis which, though based on a simple numerical example, is claimed to have some relevance. First, it is not at all clear whether Say's Law, which Ricardo adhered to so vehemently on other occasions, also operates when labour is displaced by machinery. Secondly, Ricardo simply presents the initial disruption of new machinery without saying anything about the nature of the new equilibrium or the adjustment process leading to it. This contrasts sharply with his usual emphasis upon permanent effects –indeed, in assuming that the new machines made will be able to realize 1,000 units of profit Ricardo is implicitly assuming some sort of pervasive equilibrium. Thirdly, Ricardo appears to deny the practical importance of his example at the end of the chapter when he emphasizes that his argument holds only when the new machinery is introduced suddenly.

> The statements which I have made will not, I hope, lead to the inference that machinery should not be encouraged. To elucidate the principle, I have been supposing, that improved machinery is *suddenly* discovered, and extensively used; but the truth is, that these discoveries are gradual, and rather operate in determining the employment of the capital which is saved and accumulated, than in diverting capital from its actual employment.

This point gains additional force from Ricardo's insistence that the state take no action to discourage technological progress. Most subsequent economists, from Malthus onwards, took exception to the collection of assumptions necessary to produce Ricardo's result.

Of the classical economists who followed, only Nassau Senior and John Stuart Mill tried to justify Ricardo's reasoning, sometimes with the same surprising pattern of argument that characterized Ricardo. For example, John Stuart Mill (1848) begins by asserting that workers suffer temporarily when circulating capital is converted to fixed capital; almost immediately however he adds that this is a case which scarcely ever occurs in practice. An attempt by J.E. Tozer to provide a mathematical formulation of the question does not go beyond the framework set of by Ricardo. Tozer (1838) grants that there is an initial deduction from the wages fund but points out that the fund is replenished over time as the additional output from the machinery is produced. There does not appear to be a serious effort at going beyond Ricardo's analytical schema until the writings of Knut Wicksell.

With his usual clarity, Wicksell begins his section on production and distribution by setting forth the technical conditions necessary for the validity of the marginal productivity theory of distribution. He recognizes that the distributive impact of machinery depends upon the manner in which machinery alters the marginal productivities of labour and capital and that simple answers to such a question are unlikely. One issue which he analyses with considerable acumen is the position of Ricardo that machinery may actually diminish the gross product. Wicksell takes issue with Ricardo's conclusion and claims that Ricardo did not follow his premises to their logical conclusion – under free competition, changes in technique cannot lead to a diminution of gross product. This is proved as follows:

> Let x and y denote the number of labourers per acre using the old and new methods of cultivation, and let $f(x)$ and $\varphi(y)$ denote the production functions of these lands. If m acres are cultivated by the old method and n acres by the new method, then the problem of maximizing total product is Maximize
>
> $$mf(x) + n\varphi(y)$$
>
> subject to
>
> $$m + n = B \quad mx + ny = A$$
>
> where B is the total number of acres and A is the total amount of labour available. The first order conditions for a maximum are,
>
> $$f'(x) = \varphi'(y) \text{ and } f(x) - xf'(x) = \varphi(y) - y\varphi'(y)$$

where the prime denotes differentiation. The first condition states that total product is maximized

when the marginal product of labour is equal, under both methods and the second condition states the equality of rents per acre. Wicksell now observes that these are precisely the conditions achieved by pure competition and hence that Ricardo was wrong to claim that a diminution of gross product was possible. Modern readers will note that Wicksell assumes throughout the validity of interior maxima. Subject to this qualification, Wicksell's analysis is a considerable improvement on anything produced before him. The problem just discussed considered labour and land as the only explicit factors of production. Even here, Wicksell feels that 'It is scarcely possible to discover a simple and intelligible criterion which will indicate whether a change in the technique of production is in itself likely to raise or lower wages'. When Wicksell goes on to add capital as a factor of production, he has to concede that inventions may reduce the marginal product and share of labour. This leads him to say that 'The capitalist saver is . . . fundamentally, the friend of labour, though the technical inventor is not infrequently its enemy'. (Wicksell 1911, pp. 140, 143, 164)

A satisfactory treatment of the machinery question depends upon modelling the general equilibrium of an economy and of following its transition from an initial equilibrium to the new equilibrium after the introduction of machinery. Even today, such a treatment is by no means easily achieved. Perhaps the classical economists would have done best to accept the general benefits of machinery, subject to transitional difficulties, as expounded by economists such as Tucker and Stewart, and wait until the proper analytical tools to discuss the issue satisfactorily had been developed.

See Also

- ▶ Ricardo, David (1772–1823)
- ▶ Tozer, John Edward (1806–1877)

Bibliography

Archer, H. 1742. Second parliament of George II: Fourth session (8 of 9, begins 7/4/ 1738). *The history and proceedings of the house of commons: volume 10: 1737–1739* (1742), 258–92. Available at http://www.british-history.ac.uk/report.asp?compid=37804. Accessed 19 July 2007.

Barton, J. 1817. *Observations on the circumstances which influence the conditions of the labouring classes of society.* London.

Cary, J. 1695. *A discourse on trade.* 3rd ed. London, 1745.

Defoe, D. 1704. *Giving alms no charity.* London.

Martin, H. 1701. *Considerations on the East-India Trade.* Reprinted in *A select collection of scarce and valuable economic tracts*, ed. J.R. McCulloch. London, 1856.

Mill, J.S. 1848. Principles of political economy with some of their applications to social philosophy. In *Collected works of John Stuart Mill*, ed. J.M. Robson. Toronto: University of Toronto Press. 1970.

Ricardo, D. 1821. Principles of political economy and taxation. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, 3rd ed. Cambridge: Cambridge University Press. 1951.

Stewart, D. 1856. Lectures on political economy. In *Collected works of Dugald Stewart, vols 8 and 9*, ed. Sir W. Hamilton. Edinburgh: Thomas Constable &Co. 1854–60.

Tozer, J.E. 1838. In *A mathematical investigation of the effect of machinery*, ed. D. Collard. New York: A.M. Kelley. 1968.

Tucker, J. 1757. *Instructions for travellers*. Reprinted in R.L. Schuyler, Josiah Tucker. New York: Columbia University Press. 1931.

Wicksell, K.1911. *Lectures on political economy*, 2nd ed. Trans. E. Classen. London: G. Routledge & Sons, 2 vols. 1934–5.

Machlup, Fritz (1902–1983)

John S. Chipman

Keywords

Economic integration; Haberler, G.; Innovation; International monetary system; Invention; Machlup, F.; Multiplier analysis

JEL Classifications

B31

Fritz Machlup was born in Wiener Neustadt, south of Vienna, on 15 December 1902, and died in New York on 30 January 1983. He studied economics at the University of Vienna in the 1920s under Friedrich von Wieser and Ludwig von

Mises, and wrote his doctoral dissertation on the gold-exchange standard (Machlup 1925) under the latter. In the years 1922–1932 he pursued a business career in the family cardboard-manufacturing partnership while continuing his intellectual interests in economics and philosophy of science in association with von Mises, Hayek, Haberler, Morgenstern, Felix Kaufmann and Alfred Schütz. During this period he wrote two more books including one (Machlup 1931) dealing with the role of stock-market speculation in capital formation. As business conditions deteriorated in the 1930s he took leave from his partners to accept a Rockefeller fellowship, and spent 1933–1935 in the United States. Upon receiving an appointment at the University of Buffalo in 1935 he liquidated his Austrian business interests, and following a brief stay in England began an academic career in the United States. He moved to Johns Hopkins in 1947, and to Princeton in 1960 to succeed Jacob Viner. At Hopkins he had a profound influence as a graduate teacher and in building up a first-rate graduate programme that achieved national prominence; a list of his students is contained in Machlup (1963), and tributes and testimonials from many of them will be found in Dreyer (1978). At Princeton he was extremely active in his direction of the International Finance Section. Upon his retirement in 1971 he resumed his active career at New York University until his death shortly after his 80th birthday. He was president of the Southern Economic Association (1959), the American Association of University Professors (1962–1964), the American Economic Association (1966), and the International Economic Association (1971–1974).

Machlup's two great areas of research were international monetary economics and industrial organization, the latter with special emphasis on the 'knowledge industry', an activity which began with a study of the patent system (Machlup and Penrose 1950; Machlup 1958), continued with the development of a formal theory of invention, innovation, and the optimal lag of imitation behind innovation (see Bitros 1976, pp. 439–502), and culminated in a monograph on the subject (1962), the multi-volumed second edition of which remained unfinished at the time of his death (Machlup 1980b, 1982b, 1983). What was especially original in his contributions was his peculiar talent, resulting from his business background and study of philosophy of science, of being able to formulate a theory that took into account – in addition to the usual economic facts – the theories or rationalizations put forward by economic agents to justify their own actions. This was used to support his contention that economic agents engage in maximizing behaviour even though they may deny this. Such perceptions permeate his works on industrial organization (Machlup 1949, 1952a, b) and were developed in numerous articles collected in Machlup (1963).

Machlup's contributions to international economics were likewise characterized by a combination of clear logical thinking and intimate knowledge of the workings of economic institutions. His two-country extension of the theory of the multiplier (1943) was especially illuminating in bringing out the implicit financial assumptions of Keynesian theory. His work on the theory and policy of foreign-exchange markets and international economic adjustment (collected in Machlup 1964) was very influential. His classic (1939–1940) article developing Haberler's concepts of demand and supply of foreign exchange was required reading for a generation of graduate students. In his famous controversy with Sidney Alexander, while stressing the importance of relative prices he proved himself to be always the eclectic, never espousing one narrow 'approach' to the exclusion of all others. At first countering 'elasticity pessimism' and championing flexible exchange rates in his academic writings, he later became the prime architect of plans to reform the international monetary system in his organization of the 'Bellagio group' (Machlup and Malkiel 1964). These activities have been recounted by Robert Triffin and John Williamson in Dreyer (1978). Machlup's last contributions to international economics included a series of penetrating analyses of the Eurodollar market (starting with Machlup 1970) and his one foray into 'real' international trade, a work on the theory of economic integration (Machlup 1977).

Machlup had a remarkable and unforgettable personality. He was brilliant and as sharp as a whistle in his keen analysis and grasp of economic

issues; he was lucid and patient as a teacher, yet tough; he was charming and witty; he was a great music-lover and an avid sportsman to the end of his days. Above all he was a man of extraordinary energy and passion.

Most of Machlup's important articles have been reprinted in Machlup (1963, 1964) and Bitros (1976); the first and third of these contain bibliographies of his work. Further information concerning his life and work will be found in Dreyer (1978), Chipman (1979), Machlup (1980a, 1982a), and Haberler (1983). The latter concludes with an apt poetic tribute by Kenneth Boulding.

Selected Works

1925. *Die Goldkernwährung*. Halberstadt: Meyer.

1931. *Börsenkredit, Industriekredit und Kapitalbildung*. Vienna: Julius Springer. Revised English ed., *The stock market, credit and capital formation.* London/New York: Hodge/Macmillan, 1940.

1943. *International trade and the national income multiplier*. Philadelphia: Blakiston.

1949. *The Basing-point system*. Philadelphia: Blakiston.

1950. (With E. Penrose). The patent controversy in the nineteenth century. *Journal of Economic History* 10:1–29.

1952a. *The Economics of Sellers' competition*. Baltimore: Johns Hopkins Press.

1952b. *The political economy of monopoly*. Baltimore: Johns Hopkins Press.

1958. *An economic review of the patent system*. Study of the subcommittee on patents, trademarks, and copyrights of the committee on the judiciary, US senate, study no. 15. Washington, DC: Government Printing Office.

1962. *The Production and Distribution of Knowledge in the United States*. Princeton: Princeton University Press.

1963. *Essays on economic semantics*. Englewood Cliffs: Prentice-Hall.

1964. (With G. Malkiel, eds). *International monetary arrangements: The problem of choice*. Princeton: International Finance Section, Department of Economics, Princeton University.

1964. *International payments, debts, and gold*. New York: Scribner's. 2nd ed. New York: New York University Press. 1976.

1970. Euro-dollar creation: A mystery story. *Banca Nazionale del Lavoro Quarterly Review* 23:219–260. Reprinted in Bitros (1976).

1977. *A history of thought on economic integration*. New York: Columbia University Press.

1980a. My early work on international monetary problems. *Banca Nazionale del Lavoro Quarterly Review* 35:115–46.

1980b, 1982b, 1983. *Knowledge: Its creation, distribution, and economic significance*. Vol. 1: *Knowledge and knowledge production*. Vol. 2: *The branches of learning*. Vol. 3: *The economics of information and human capital*. Princeton: Princeton University Press.

1982a. My work on international monetary problems, 1940–1964. *Banca Nazionale del Lavoro Quarterly Review* 35, 3–36.

Bibliography

Bitros, G., ed. 1976. *Selected Economic Writings of Fritz Machlup*. New York: New York University Press.

Chipman, J.S. 1979. Machlup, Fritz. In *International encyclopedia of the social sciences*, vol. 18. New York: Free Press.

Dreyer, J.S., ed. 1978. *Breadth and depth in economics: Fritz Machlup – the man and his ideas*. Lexington: D.C. Heath.

Haberler, G. 1983. Fritz Machlup zum Gedenken. Neue Zürcher Zeitung, 16 February, 19. Fritz Machlup: in memoriam. Cato Journal 3: 11–14.

Macleod, Henry Dunning (1821–1902)

Murray Milgate and Alastair Levy

Keywords

Banking system; Bimetallism; Catallactics; Gresham's Law; Macleod, H. D.

JEL Classifications

B31

Macleod warrants special mention in this *Dictionary* if only because in the late 1850s and early 1860s he undertook to produce single-handedly a dictionary of economics on a grand scale – and, what is more, one to which he was to be the sole contributor. In the event the task proved to be beyond him, as it would for any mortal, and all that appeared was the first volume covering the letters A–C. Macleod never held an academic appointment, though he applied unsuccessfully for chairs at Cambridge (1863), Edinburgh (1871), and Oxford (1888).

Macleod, the son of a Scottish landholder, was born in Edinburgh. After graduation from Cambridge (BA, Trinity, 1843) and admission to the Bar (1849), he wrote a report on the administration of poor relief in the nine local parishes of the district of Easter Ross in Scotland (1851). This report led directly to the establishment of a poor-house under Scotland's first Poor Law Union. In 1854, he joined the Royal British Bank and wrote a memorandum and opinion on that bank's legal position under the Joint Stock Banking Act of 1845. This first excursion into financial matters stimulated him to study the literature of economics on the subject, but he found that

> for the purpose of describing the actual principles and mechanisms of commerce they [Smith, Ricardo and Mill] were absolutely worthless. . . . I saw that the greatest opportunity that had come to any man since Galileo had come to me, and I then determined to devote myself to the construction of a real science of Economics on the model of the already established physical sciences. (1896b, pp. 142–3)

To his credit he stuck fast to his task. His detractors, however, have passed harsh judgement on its results (see, for example, the assessment of him in Higgs's edition of *Palgrave's Dictionary*); his sympathetic readers have been more generous (see, for example, Hayek 1933). Given the sheer magnitude of his project, there would seem to be more to be said for the position of the critics.

Macleod's employment in the banking system led to what is perhaps his most important book on that subject: *The Theory and Practice of Banking* (1855–6). Its two most interesting features for the modern reader are, perhaps, the discussion of discount policy and the insistence on the proposition that 'the distinction between capital and currency . . . is of the most profound delusions that ever existed' (1855, vol. 2, p. lxxii; see also the entry on 'Credit' in his *Dictionary).* Not surprisingly, for one who kept fast to the basic position of the Bullion Report, this latter notion introduced a number of ambiguities into the argument. However, not withstanding these peculiarities, the book was apparently quite successful, going through five editions by the 1890s and being reprinted soon after his death. Charles Rist referred to it as Macleod's 'great book' (1940, p. 261).

There followed many publications on monetary matters of which two may be singled out. In *Bimetallism* (1894), he criticized the proponents of a dual standard for advocating 'an impossibility'; a position which put him at odds with many of his contemporaries. This polemic was continued in two short tracts issued by the Gold Standard Defence Association in 1895 under the titles 'Gresham's Law' and 'Bimetallism in France'. Secondly, in 1898, he published two contributions to the debate surrounding the Fowler Commission on Indian currency arrangements: *Indian Currency* and *A Tentative Scheme for Restoring a Gold Currency to India.*

Macleod's project of reconstructing economic science continued on more general matters with *Elements of Political Economy* (1858). The book is interesting as an example of Macleod's advocacy of a definition of economics as the 'science of exchanges', or catallactics, which Marshall claimed 'anticipated much both of the form and substance of recent criticisms on the classical doctrine of value in relation to cost, by Profs. Walras and Carl Menger' (1920, p. 821), and for the fact that in it he introduces into the vocabulary of economics the phrase 'Gresham's Law'.

One of Macleod's interesting habits was that of publishing the same material in different forms and under different titles. In this he reminds one of McCulloch. Thus, *Bimetallism* was itself an expanded version of the seventh chapter of his *Theory of Credit* (1889–91), his *Elements of Political Economy* (1858) appeared in successive editions under the titles *The Principles of Economic*

Philosophy (1872–5) and *The Elements of Economics* (1881–6), and his *History of Economics* (1896b) seems to be made up of material from his unfinished dictionary.

Macleod died at Norwood on 16 July 1902.

See Also

► Catallactics

Selected Works

1851. *The results of the operation of the poor-house system in ross*. Inverness: Courier Office, repr. from the Inverness Courier.
1855–6. *The theory and practice of banking; with the elementary principles of currency, prices, credit and exchanges,* 2 vols. London: Longman, Brown, Green and Longmans.
1858. *The elements of political economy.* London: Longman, Brown, Green, Longmans and Roberts.
1863. *A dictionary of political economy: Biographical, bibliographical, historical and practical: A–C.* London: Longmans, Green, Longman, Roberts, and Green.
1889–91. *The theory of credit*, 2 vols. London: Longmans, Green & Co.
1894. *Bimetallism.* London: Longmans, Green & Co.
1896a. *A history of banking in all the leading nations,* 2 vols. London: Bliss, Sands & Co.
1896b. *The history of economics.* London: Bliss, Sands & Co.
1898. *Indian currency.* Longmans: Green & Co.

Bibliography

Hayek, F.A. 1933. Henry Dunning Macleod. In *Encyclopedia of the social sciences*, ed. E.R.A. Seligman, vol. 10. New York: Macmillan.
Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.
Rist, C. 1940. *History of monetary and credit theory from John law to the present day.* Trans. J. Degras from the French ed of 1938, New York/London: Macmillan/G. Allen & Unwin. Reprinted, New York: A.M. Kelley, 1966.

Macroeconometric Models

Ray C. Fair

The topic 'macroeconometric models' is very broad. Conceivably it could include any study in which at least one equation was estimated using macroeconomic data. I will limit my discussion to structural models that try to explain the overall economy, although I will also have a few things to say about vector autoregressive models.

One might have thought at the beginning of large-scale model construction in the early 1950s that there would be a gradual and fairly systematic improvement in the accuracy of the specification of the equations, so that by the late 1980s there would exist a generally agreed upon model. This is not the case. Macroeconomics has now been in a state of flux for more than a decade. There is currently much disagreement among macroeconomists about the structure of the economy, and there is certainly no generally agreed upon model. This lack of agreement manifests itself in quite different monetary and fiscal policy recommendations that are made at any one time by different economists.

The unsettled nature of macroeconomics makes it an exciting research area, but also a difficult one to review in a short essay. Any current review of macroeconometric models must be selective and somewhat idiosyncratic, and this is certainly true of the present review. I have chosen some topics that I think are important in the area, but the list is by no means exhaustive.

A Brief History

A comprehensive discussion of the history of macroeconometric model-building is in Bodkin et al. (1986). The following discussion is very brief. The beginning of the construction of macroeconometric models is usually traced back to Tinbergen's (1939) work on business cycles in the late 1930s, although there were earlier efforts

by Tinbergen and others that could be classified as macroeconometric models. Tinbergen's model was an annual model and consisted of 31 behavioural equations and 17 identities. It was estimated by ordinary least squares for the period 1919–32. There are few exogenous variables in the model; Bodkin et al. (1986) have identified only five. Tinbergen never 'solved' his model in the modern sense. He did, however, spend considerable time analysing the dynamic properties of the model analytically after reducing the model to a linear difference equation in corporate profits.

Work on macroeconometric models began in earnest after World War II. The leading figure in the postwar period has been Lawrence Klein, who built his first models while at the Cowles Commission in the mid-1940s. The results of this work were published in Klein (1950). This monograph contained three models. Model I contains three behavioural equations and three identities, and it was estimated by full-information maximum likelihood, limited-information maximum likelihood and ordinary least squares. This model has been an important pedagogical tool for many decades. Model II contains a consumption function and two identities. Model III contains 12 behavioural equations and 4 identities. It is best known today as the precursor of the Klein–Goldberger model.

The Klein–Goldberger model (1955) began as a project of the Research Seminar in Quantitative Economics at the University of Michigan. It was an annual model, estimated for the split-sample period 1929–41, 1946–52. It consisted of 15 behavioural equations and 5 identities and was estimated by limited-information maximum likelihood. It was the first model to be used for regular *ex ante* forecasting purposes. The first forecast was for the year 1953.

The 1960s was an active time for model-builders. One of the major efforts of the decade was the Brookings model, a quarterly model which at its peak contained nearly 400 equations. This model was a joint effort of many individuals, and although it never achieved the success that was initially expected, much was learned during the effort. The model was laid to rest around 1972.

The 1970s was a time in which many models became commercially successful, the most successful being DRI (Data Resources, Inc.), Wharton, and Chase. The commercialization of the models changed the focus of research somewhat. Less time was spent on what one might call basic research on the models, such as experimenting with alternative estimation techniques and statistical testing. More time was spent on the month-to-month needs of keeping the models up to date. The 1970s was also a time in which macroeconometric models came under attack on academic grounds. The key attack was the Lucas (1976) critique, which argued that the models are not likely to be useful for policy purposes. More will be said about this later. These attacks had the effect of moving academics away from research on large-scale models to other things, and very little academic research was done on models in the 1970s. This has continued to be true to some extent in the first half of the 1980s.

Methodology

Economic theory is used to guide the specification of macroeconometric models. The 'traditional' procedure is to use theory to guide the choice of explanatory variables in the equations to be estimated. Consider, for example, the multi-period utility-maximization theory of household behaviour. This theory tells us that a household's consumption and labour-supply decisions are a function of the prices of the goods, the wage rate, non-labour income, interest rates and the initial value of wealth. The traditional procedure is thus to use these as explanatory variables in the equations explaining consumption and labour supply. In many cases theory indicates the signs of the coefficients of the explanatory variables. Theory is generally not used to decide the functional forms of the estimated equations and the lengths of the lag distributions. Functional forms and lag lengths are generally chosen empirically, by trying alternative forms and lengths to see which produces the best results.

The transition from theoretical models to empirical models is a difficult problem

in macroeconomics. One is usually severely constrained by the quantity and quality of the available data, and many restrictive assumptions are generally needed in the transition from the theory to the data. In other words, extra 'theorizing' occurs during the transition, and it is usually theory that is less appealing than that of the purely theoretical model.

The place where extra theorizing occurs most is the treatment of expectations. Expected future values play an important role in most theoretical models. In the multi-period utility-maximization model, for example, expected future values of prices, the wage rate, non-labour income and interest rates affect current consumption and labour-supply decisions. Since expected values are generally not observed, one needs to make some assumption about how expectations are formed when specifying the empirical equations to estimate. A common approach is to assume that expectations of a particular variable are a function of current and lagged values of the variable. Under this assumption one simply replaces the expected future values with current and lagged values. This assumption is fairly ad hoc, and much of the research in macroeconomics in the last decade is on the question of how expectations are formed. More will be said about expectations later.

Once enough assumptions have been made so that only observed variables appear in the equations, the equations are ready to be estimated. The estimation techniques range from ordinary least squares and two-stage least squares to three-stage least squares and full information maximum likelihood. Many equations in macroeconometric models have serially correlated error terms, and a common procedure is to estimate the equations under the assumption that the error terms are first-order autoregressive. If the model is simultaneous, which almost all models are, ordinary least squares produces inconsistent estimates.

Much experimentation takes place at the estimation stage. Different functional forms and lag lengths are tried, and explanatory variables are dropped if they have coefficient estimates of the wrong expected sign. Variables with coefficient estimates of the right sign may also be dropped if the estimates have t-statistics that are less than about two in absolute value, although practice varies on this. If things are not working out very well in the sense that very few significant estimates of the correct sign are being obtained, one may go back and rethink the theory or the transition from the theory to the estimated equations. This process may lead to new equations to try and perhaps to better results. This back-and-forth movement between theory and results can be an important part of the construction of the model.

The estimation technique that is used in experimenting with alternative specification is usually a limited-information technique, such as two-stage least squares. These techniques have the advantage that one can experiment with a particular equation without worrying very much about the other equations in the model. Knowledge of the general features of the other equations is used in the choice of the first-stage regressors for the two-stages least squares technique, for example, but one does not need to know the exact features of each equation when making this choice. If a full-information technique is used, it is usually used at the end of the search process to estimate the final version of the model. If the full-information estimates are quite different from the limited-information ones, it may again be necessary to go back and rethink the theory and transition. In particular, this may indicate that the version of the model that has been chosen by the limited-information searching is seriously misspecified. Sometimes ordinary least squares is used in the searching process even though the model is simultaneous. This, however, has little to recommend it since the ordinary least squares estimates are inconsistent, and consistent alternatives like two-stage least squares are not expensive to use.

The next step after the model has been estimated is to test and analyse it. One way in which models are tested is to compute predicted values from solving the overall model and compare the predicted values to the actual values. The accuracy of the predictions is usually examined by calculating room mean-squared errors. The properties of models are analysed by performing 'multiplier' experiments. These experiments involve

changing one or more exogeneous variables and observing how the predicted values of the endogenous variables change. Models can also be analysed by performing optimal-control experiments. Given a particular objective function and given a set of policy variables, one can find the values of the policy variables that maximize the objective function subject to the constraints imposed by the model.

It may also be the case that things are not working out well at the testing and analysis stage. Poor fits may be obtained; multipliers that seem (according to one's *a priori* views) too large or too small may be obtained; and optimal control experiments may yield optimal values that do not seem sensible. This may also lead one to rethink the theory, the transition to the estimated equations, or both, and perhaps to try alternative specifications. The back-and-forth movement between theory and results may thus occur at both the estimation and analysis steps.

It is important to note that the back-and-forth movement between theory and results may yield a model that fits the data well and seems on other grounds to be quite good, when it is in fact a poor approximation of the structure of the economy. If one searches hard enough, it is usually possible with macro time-series data to come up with what seems to be a good model. The searching for models in this way is sometimes called 'data mining' and sometimes called 'specification searches', depending on one's mood. Fortunately, there is a way of testing whether one has mined the data in an inappropriate way, which is to do outside sample tests. If a model is poorly specified, it should not fit well outside the sample period for which it was estimated, even though it looks good within sample. It is thus possible to test for misspecification by examining outside sample results. A method for doing this is discussed in the next section.

Because of the dropping of variables with wrong signs and (possibly) the back-and-forth movement from multiplier results to theory, an econometric model is likely to have multiplier properties that are similar to what one expects from the theory. Therefore, the fact that an econometric model has properties that are consistent with the theory is in no way a confirmation of the model, at least in my view. Models must be tested by using methods like the one discussed in the next section, not by examining the 'reasonableness' of their multiplier properties.

There are two main alternatives to the traditional procedure just outlined. One, which is discussed in section "The Lucas Critique and the Estimation of Deep Structural Parameters", is to take more seriously the theoretical restrictions that are implied by the assumption that decisions are made by maximizing objective functions. The other alternative is to estimate vector autoregressive models, where very few theoretical restrictions are imposed. This alternative has been stressed by Sims (1980). A vector autoregressive (VAR) model is one in which each variable is regressed on lagged values of itself and lagged values of all the other variables in the model. This approach imposes some restrictions on the data – in particular, the number of variables to use, the lengths of the lags and (sometimes) cross-equation restrictions on the coefficients – but the restrictions are in general less restrictive than the exclusionary ones used by the traditional approach. As discussed in the next section, VAR models are useful for comparison purposes even if one otherwise does not agree with the VAR methodology.

Macroeconometric models are also used to make forecasts. Given a set of coefficient estimates, a set of values of the future-error terms, and a set of guesses of the future values of the exogenous variables, one can use a model to make predictions of the future values of the endogenous variables. A forecast beyond the data, where guessed values of the exogenous variables must be used, is called an *ex ante* forecast. A forecast within the data, where actual values of the exogenous variables are used, is called an *ex post* forecast. The values chosen for the error terms are usually the expected values, which are almost always zero. If an equation has been estimated under the assumption of a first-order autoregressive error, the estimate of the autoregressive coefficient and last period's estimated-error term are used in estimating the current period's error term.

In practice, *ex ante* forecasts are often 'subjectively adjusted'. If, when unadjusted, the model is not forecasting what the model-builder thinks is

going to happen, the equations are changed by adding or subtracting values from the constant terms. In many cases a constant term in an equation is changed more than once over the forecast horizon. Adjusting the values of constant terms is equivalent to adjusting the values of the error terms, given that a different value of the constant term can be used each period. This procedure can thus be looked on as the model-builder guessing the future values of the error terms. Instead of setting the future-error terms equal to their expected values, the model-builder overrides this aspect of the model and sets values based on his or her own feelings about what is going to happen. With enough adjustments it is possible to have the forecasts be whatever the user wants, subject to the restriction that the identities must be satisfied. This means, of course, that in practice one can never be sure how much of the forecast is due to the model and how much is due to the model-builder. It also means that *ex ante* forecasts are of little use for testing and comparing models *qua* models.

Testing

The testing of macroeconometric models is extremely difficult, and this is undoubtedly one of the reasons that there is so little agreement in the area. There are two main problems in comparing different models. First, models may differ in the number and types of variables that are taken to be exogenous. If, for example, one model takes prices as exogenous whereas a second model does not, the first model has an obvious advantage over the second in predictive tests. Second, data mining may make a model look good within sample when it is in fact a poor approximation of the structure.

I have developed a method for comparing models that helps account for these problems (Fair 1980). The method is briefly as follows. There are four main sources of uncertainty of a forecast from a model: uncertainty due to (1) the error terms, (2) the coefficient estimates, (3) the exogenous variables and (4) the possible misspecification of the model. Uncertainty from the error terms and coefficient estimates can be estimated using stochastic simulation. From the estimation of the model one has estimates of the covariance matrix of the error terms and the covariance matrix of the coefficient estimates. Given these estimates and given an assumption about the functional form of the distributions, such as normality, one can draw error terms and coefficients. For a given set of draws the model can be solved and the predicted values of the endogenous variables recorded. This is one trial. Many trials can be performed, and after, say, J trials, one has J predicted values of each endogenous variable for each period of the forecast. From the J values one can compute the mean of the forecast and the variance of the forecast error.

In order to account for exogenous-variable uncertainty, one needs some assumption about the uncertainty itself. One polar assumption is that the exogenous variables are in some sense as uncertain as the endogenous variables. One might, for example, estimate autoregressive equations for each exogenous variable and add these equations to the model. This would produce a model with no exogenous variables, which could then be tested. The other polar assumption is that there is no uncertainty attached to the exogenous variables. This might be true of some policy variables. I have generally worked with an in-between case, where I estimate an autoregressive equation for each exogenous variable and use the estimated variance from this equation as an estimate of the variance of the exogenous-variable forecast error. I the use these estimated variances and the normality assumption to draw values for the exogenous variables. Each trial of the stochastic simulation thus consists of draws of the error terms, coefficients and exogenous variables.

Estimating the degree of misspecification of the model is based on a comparison of two estimated forecast-error variances. One is the stochastic simulation estimate; the other is the square of the outside-sample forecast error. If the model is correctly specified, the expected value of the difference between the two estimates is zero (ignoring simulation error). If one has data mined in an inappropriate way and the model is misspecified, one would expect the stochastic simulation estimate to be smaller than the estimate

from the outside-sample error. The expected value of the different is thus likely to be positive for a misspecified model. By repeated re-estimation and stochastic simulation of the model, where one observation is added at the end of the sample period for each estimation, the expected value of the difference between the estimated variances can be estimated. The differences are then estimates of the degree of misspecification of the model.

The final estimated forecast-error variance for each variable for each period is obtained by adding the estimated difference to the stochastic-simulation estimate that is based on draws of the error terms, coefficients and exogenous variables. This estimated variance has accounted for the four main sources of uncertainty of a forecast, and it can be compared across models. Speaking loosely, each model is on an equal footing for comparison purposes. If one model has smaller estimated variances than another, this is evidence in favour of the model. Autoregressive and vector autoregressive models are useful for comparison purposes. Estimated variances of a structural model can be compared to those of an autoregressive or vector autoregressive model. If the estimated variances of the structural model are in general larger after taking all the sources of uncertainty into account, this is a cause of some concern.

The Lucas Critique and the Estimation of Deep Structural Parameters

The theory that is used to guide the specification of econometric equations in what I am calling the traditional approach is generally based on some implicit objective function that is being maximized. The parameters of the objective function are not, however, directly estimated. The parameters of the derived-decision equations are estimated instead, where the estimated parameters are combinations of the parameters of the objective function, the parameters of expectation-formation mechanisms, and other things. A problem with estimating combinations is that if, say, one wants to examine the effects of changing an exogenous variable or a policy rule on the decision variables, there is always the possibility that this change will change something in the combinations. If so, then it is inappropriate to use the estimated-decision equations, which are based on fixed estimates of the combinations, to examine the effects of the change. This is the point emphasized by Lucas (1976) in his critique of macroeconometric models.

Lucas's critique has led to a line of research concerned with estimating parameters of objective functions, which are sometimes called 'deep' structural parameters. These parameters, which are primarily taste and technology parameters, are assumed not to change when policy rules and exogenous variables change, and so one can use them and the associated model to examine the effects of policy changes. The approach is appealing in this sense, although many restrictive assumptions are involved in setting up the estimation problem, such as the specification of a particular form for the objective function. It is too early to know how useful this approach will be in practice.

If the approach of estimating deep structural parameters turns out not to lead to econometric models that are good approximations, this does not invalidate Lucas's critique. The critique is a logical one. If parameters that are taken to be constant change when policy changes, the estimated effects of the change are clearly in error. The key question for any experiment with a model is the likely size of the error. There are many potential sources of error, and even the best econometric model is only an approximation. One of the most important sources of error in my view is the use of aggregate data. As the age and income distributions of the population change, the coefficients in aggregate equations are likely to change, and this is a source of error in the estimated equations. This problem may be quantitatively much more important than the problem raised by Lucas.

One encouraging feature regarding the Lucas critique is the following. Assume that for an equation or set of equations the parameters change considerably when a given policy variable changes. Assume also that the policy variable changes frequently. In this case the method discussed in section "Testing" is likely to reject a

model that includes this equation or set of equations. The model is obviously misspecified, and the method should be able to pick up this misspecification if there have been frequent changes in the policy variable. One may, or course, still be misled regarding the Lucas critique if the policy variable has changed not at all or very little in the past. In this case the model will still be misspecified, but the misspecification has not been given a change to be picked up in the data. One should thus be wary of drawing conclusions about the effects of seldom-changed policy variables unless one has strong reasons for believing that the Lucas critique is not quantitatively important for the particular policy variable in question.

Models with Rational Expectations

In the past few years research has begun on macroeconometric models with rational expectations. Consider a model in which some of the explanatory variables are expected future values. In particular, assume that y^e_{t+1} appears as an explanatory variable in the first equation, where y^e_{t+1} is the expected value of y for period $t + 1$ based on information through period $t - 1$. In the utility-maximization model in section "A Brief History", the equation being estimated might be a consumption equation and y might be the wage rate. If expectations are assumed to be rational in the sense of Muth (1961), then the value of y^e_{t+1} is equal to the model's prediction of y for period $t + 1$. In other words, the expectation of a variable is equal to the model's prediction of it. Under the assumption of rational expectations, agents know the model and use it to generate their expectations. Agents are obviously assumed to be much more sophisticated in this case than they are in the case in which expectations of a variable are simply a function of current and lagged values of the variable.

Models with rational expectations are more difficult to estimate and solve than are standard models. Two types of estimation methods have been proposed for these models. One is full-information maximum likelihood, FIML (Fair and Taylor 1983). This method accounts for all the restrictions that are implied by the rational-expectations hypothesis, including all cross-equation restrictions. Unfortunately, the method is expensive to use, and it is not currently computationally feasible for large non-linear models.

There are limited information alternatives to FIML. The main alternative is Hansen's (1982) method of moments estimator. Limited-information methods like Hansen's estimator are based on the assumption that agents form expectations rationally and that there is an observed vector of variables (observed by the econometrician), denoted Z_t that is used in part by agents in forming their (rational) expectations. The methods do not require for consistent estimates that Z_t include all the variables used by agents in forming their expectations. Limited-information techniques are not very expensive to compute, and they have been widely used in practice.

The solution of rational expectations models is more difficult than the solution of standard models because future predicted values affect present predicted values. In other words, one cannot solve for the present without knowing the future. The solution method that has come to be used (Fair and Taylor 1983) iterates on solution *paths*. One guesses paths for the future values of the expectations and then solves the model period by period, treating the paths as predetermined. This solution yields new paths for the future values of the expectations, and so the model can be solved again period by period, treating the new paths as predetermined. This then yields new solution paths, which can be used for a new period-by-period solution, and so on. Convergence is achieved when the solution paths on one iteration are within some prescribed tolerance level of the solution paths on the next iteration. This method turns out to work quite well in practice and is not that expensive.

Work is essentially just beginning on macroeconometric models with rational expectations, and no strong conclusions can as yet be drawn. Results are presented in Fair (1985) that provide only mild support for the use of more sophisticated expectational hypotheses than are traditionally used in model-building. More work, however, is clearly needed. It should be noted

finally that macroeconometric models with rational expectations in the sense described here do not necessarily satisfy the Lucas critique. Depending on the set up, the coefficients that are estimated in the stochastic equations are not necessarily deep structural parameters even if there are expected future variables among the explanatory variables.

Conclusion

Work in macroeconometrics has the advantage that new observations are continually being generated. The current range of disagreement in macroeconomics may be narrowed in the future as more data are generated and more tests performed. Whether this will happen and whether there will be a return to more academic research on macroeconometric models is hard to say. Academic research on models clearly peaked in the 1960s, and it may have reached a trough in the late 1970s or early 1980s. But trying to predict research cycles is probably more hazardous than trying to predict business cycles.

See Also

- ▶ Autoregressive and Moving-Average Time-Series Processes
- ▶ Estimation
- ▶ Rational Expectations
- ▶ Simultaneous Equations Models
- ▶ Specification Problems in Econometrics
- ▶ Tinbergen, Jan (1903–1994)

Bibliography

Bodkin, R.G., Klein, L.R., and Marwah, K. 1986. *A history of macro-econometric model-building*. Forthcoming.

Fair, R.C. 1980. Estimating the expected predictive accuracy of econometric models. *International Economic Review* 21: 355–378.

Fair, R.C. 1985. *The use of expected future variables in macroeconometric models*. Mimeo.

Fair, R.C., and J. Taylor. 1983. Solution and maximum likelihood estimation of dynamic rational expectations models. *Econometrica* 51: 1169–1185.

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

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

Klein, L.R., and A.S. Goldberger. 1955. *An econometric model of the United States 1929–1952*. Amsterdam: North-Holland.

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

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.

Tinbergen, J. 1939. *Statistical testing of business cycle theories*. Geneva: League of Nations.

Macroeconomic Effects of International Trade

Reuven Glick

M

Abstract

International trade can affect the macroeconomy by helping to transmit disturbances from one economy to another and by muting or amplifying the impact of fiscal and monetary policies on economic activity. Representative open economy macro models are discussed, highlighting the role different theoretical features play in influencing the channels through which trade flows can have macro effects.

Keywords

Balassa–Samuelson effect; Beggar-thy-neighbour; Capital mobility; Comparative advantage; Consumption correlations puzzle; Contagion; Currency unions; Elasticities approach to the balance of payments; Elasticity of substitution; Endogenous growth models; Exchange rate puzzles; Exchange rate regimes; Externalities; Feldstein–Horioka puzzle; Flexible exchange rates; General equilibrium; Imperfect competition; Income–expenditure

models; Income multiplier; Inflation; International real business cycles; International trade; Keynesian model; Macroeconomic effects of international trade; Marshall–Lerner–Robinson condition; Microfoundations; Monopolistic competition; Mundell–Fleming model; National income; New openeconomy macroeconomics; Production sharing; Purchasing power parity; Research and development; Specialization; Spillovers; Technological progress; Terms of trade; Trade costs; Trade frictions

JEL Classifications
F4

The field of open economy macroeconomics deals with the macroeconomic behaviour of economies that trade with each other. International trade can have macroeconomic effects by helping the transmission of disturbances from one economy to another as well as by affecting the impact of macroeconomic policies on economic activity. This article discusses several representative open economy macro models, highlighting the role different theoretical features play in influencing the channels through which trade flows can have macro effects.

Keynesian Framework

At its simplest level, international trade is linked to macroeconomic activity through the national income relation. Consider the Keynesian income–expenditure model of a small open economy, in which prices and the interest rate are given, foreign demand for exports is exogenous, and domestic output is determined by demand. With these assumptions, an exogenous increase in domestic expenditures raises domestic income and worsens the current account balance; however, income rises less than in a closed economy because of leakages from the income stream through imports and through saving. In contrast, an exogenous increase in foreign demand for domestic goods leads to an increase in both exports and domestic income. Because the increased direct demand for exports is only partially offset by the expansion of imports induced by higher income, the current account improves overall. The resulting rise in domestic output implies positive cross-country transmission of the foreign disturbance.

Income multiplier effects through changes in trade also characterize open economy extensions of the Keynesian framework, such as the classic Mundell–Fleming model. This model also takes prices as given, but allows the income effects of monetary stimulus and exogenous expenditure changes to take account of interest rate changes depending on the degree of international capital mobility and of exchange rate changes, which in turn depend on the exchange rate regime. With a flexible exchange rate regime, exchange rate changes affect the relative demand for domestic and foreign goods. Thus, for example, domestic monetary stimulus that reduces the interest rate, raises income, and creates an excess demand for foreign exchange also depreciates the domestic currency. If the Marshall–Lerner–Robinson condition is satisfied, that is, the sum of price elasticities of domestic and foreign demands for imports exceeds unity, then the lower relative price of domestic goods switches demand from foreign to domestic goods and raises the current account balance, causing domestic income to increase and foreign income to decrease. Accordingly, the domestic income multiplier effect of the monetary stimulus is augmented by the expenditure-switching effect of the exchange rate; in addition, the trade transmission effect of domestic monetary shocks to foreign income is negative.

In these models crucial parameters affecting transmission effects include the marginal propensity to import and the elasticity of trade with respect to the exchange rate. Thus, for example, an increase in the marginal propensity to import out of income lessens the multiplier effects of domestic policy stimulus.

New Open Economy Macro Models

New open-economy macroeconomic models (NOEM) integrate older fixed-price Keynesian

models of macroeconomic fluctuations with dynamic intertemporal analysis based on microeconomic foundations and optimizing agents. These models embed imperfect competition and short-run nominal rigidities in a general equilibrium framework and provide clear welfare criteria in the form of the utility of the representative consumer. They also assume that bond (but not equity) markets are integrated, providing a consumption-smoothing role for net trade flows via the current account. Thus, for example, a temporary productivity shock that raises domestic output induces higher saving and a temporary current account surplus (though with investment dynamics a current account deficit may result if the increase in investment exceeds the increase in saving).

In a seminal paper, Obstfeld and Rogoff (1995) use a two-country framework in which each country specializes in producing a subset of tradable goods, and domestic and foreign consumers have identical preferences over a basket of both domestic and foreign goods. They show that monetary shocks have a positive effect on domestic output and a negative transmission effect on foreign output, as in the Mundell–Fleming model. Because monetary stimulus depreciates the domestic currency, it lowers the domestic country's terms of trade, reduces the purchasing power of domestic residents and raises the purchasing power of foreign residents. This terms-of-trade effect makes foreign residents better off and domestic residents worse off, but not by enough to offset the domestic gains from greater output. A temporary current account surplus is generated as well via the intertemporal consumption-smoothing channel.

A key parameter in NOEM models is the elasticity of substitution between goods embedded in consumer preferences. Obstfeld and Rogoff assume that the elasticity of substitution between goods produced in the same country is the same as the elasticity of substitution between goods produced in different countries. Several papers show how the international transmission of shocks is affected by relaxing this assumption. Tille (2001) shows that, if the elasticity of substitution of domestic and foreign goods exceeds unity, the Marshall–Lerner–Robinson condition holds. In this case, a currency depreciation and decline in the terms of trade results in a large demand switch towards domestic goods and a rise in export revenue. Tille also shows that, if there is less substitutability between domestic and foreign goods across countries than within countries (the empirically more relevant case), the terms-oftrade effect of domestic monetary expansion may be large enough to lower domestic welfare (termed a 'beggar-thyself' effect), while raising foreign welfare. In contrast, greater fiscal expenditures on domestic output raise the domestic terms of trade and domestic welfare, while reducing relative demand for foreign goods and foreign welfare (a 'beggar-thy-neighbour' effect), particularly when domestic and foreign goods are poor substitutes.

Corsetti and Pesenti (2001) deal with the special case in which the elasticity of substitution between domestic and foreign goods is unity, implying constant expenditure shares on domestic and foreign goods. This specification implies that the current account is always in balance. The reason is that, with unit elasticity between domestic and foreign goods, an increase in the foreign price of foreign goods results in a proportionate decrease in the quantity of foreign demand for domestic goods, leaving expenditures on exports constant and the current account unaffected.

Other extensions to NOEM models that affect the transmission of policy include consumption bias for domestic over foreign goods (Warnock 2003), pricing-tomarket behaviour (Betts and Devereux 1998), and non-traded distribution services (Burstein et al. 2006).

International Real Business Cycle Models

The tendency of macro aggregates, such as output, to move together in different countries is well documented (Backus et al. 1992; Baxter 1995). Cross-country business cycle correlations depend on the interaction of common international shocks, country-specific shocks, and the transmission of these shocks between countries. An important question in international macroeconomics is how much these comovements reflect the transmission of shocks across borders through

international trade linkages. International real business cycle (IRBC) models analyse this issue within a dynamic general equilibrium framework based on microfoundations. Unlike NOEM models, these models typically assume flexible prices and complete markets, though more recent work has introduced price rigidity and incomplete asset markets.

On theoretical grounds, the effect of international trade links on the comovement of national business cycles is ambiguous. On the one hand, greater integration can increase intra-industry specialization and production-sharing because of low elasticity of substitution between intermediate inputs produced in different countries; in addition, it may allow demand shocks to propagate more easily across national borders, which may lead to a higher correlation of business cycles when countries trade more. On the other hand, greater trade integration can increase interindustry specialization if countries specialize more in the goods in which they have a comparative advantage in order to achieve gains from trade; this case, if industryspecific shocks are a dominant source of business cycle movements, may lead to a lower correlation of business cycles when countries trade more.

On balance, the empirical evidence suggests that the former effect dominates, and that countries with a lot of bilateral trade tend to have more synchronized business cycles (for example, Frankel and Rose 1998; Baxter and Kouparitsas 2005). However, since the early 1980s business cycle synchronization has not in fact increased among industrial countries despite increasing trade integration. Stock and Watson (2005) provide a partial explanation by showing that common international shocks experienced by G-7 countries have been smaller in the 1980s and 1990s than they were in the 1960s and 1970s. But they also show that cyclical comovements have increased for subgroups of countries, notably within Europe and North America. Burstein et al. (2005) construct a model that is consistent with this development in which trade between core countries and their periphery (for example, the United States and Canada) involves more production sharing than does trade between core regions (for example, the United States and Europe). Consequently, one should observe higher output correlations between core and peripheral countries than between core regions. IRBC models have been less successful in explaining the quantitative magnitude of the relation between trade intensity and the cross-country correlation of business cycles; that is, a given change in bilateral trade intensity generates a much smaller change in output correlations than is apparent in the data; this is referred to as the 'trade comovement gap puzzle' (Kose and Yi 2006).

The finding that greater trade intensity is associated with greater cross-country comovements in business cycles suggests that these comovements depend on policies that enhance international trade, such as lowering of trade barriers or reductions in exchange rate costs due to membership in currency unions. Frankel and Rose (2002) find that the positive effect of currency unions on trade in turn has a large effect on output in member countries. Since the main cost of joining a currency area is the cost of giving up monetary independence, this has the implication that a pair of countries with business cycles that are dissimilar *ex ante* (making the act of joining a currency union appear costly) might have more correlated business cycles *ex post* because the increase in trade stimulated by the currency union tends to synchronize business cycles.

Trade Frictions and Macro Models

The international tradability of goods depends not just on the degree of substitutability in consumption, but also on transport costs and other trade frictions. In fact, Obstfeld and Rogoff (2000) argue that introducing real trade costs helps explain a variety of puzzles in international economics, including the low crosscountry correlation of consumption (consumption correlations puzzle), the limited magnitude of current account imbalances (Feldstein–Horioka puzzle), international price discrepancies (purchasing power parity puzzle), and home bias in trade and asset holdings.

Taken to the extreme, trade frictions play a role in explaining why some goods may not be traded at all. While open economy macroeconomics by

definition analyses trade across national borders, the field has long found it useful to assume that a given exogenous set of goods is non-traded. This traded/non-traded distinction is essential to many well-known results in the field, such as the Balassa–Samuelson effect, which says that, as the productivity of traded goods rises relative to that of non-traded goods, there will be tendency for the real exchange rate to appreciate.

The international trade literature has explained non-tradedness as an outcome of trade frictions. For example, Dornbusch et al. (1977) show how a range of non-traded goods can arise in the presence of cross-country trade costs within a model in which differences in labour productivity across a continuum of goods determine the range of goods a country produces as well as the pattern of trade.

A growing field of international economics research tries to integrate models of trade and macroeconomics and treats the set of tradable goods not as exogenously given but rather as an endogenously determined characteristic of the analysis. Several authors (Ghironi and Melitz 2005; Bergin et al. 2006) formulate open economy macro models with monopolistic competition and heterogeneously productive firms, in which firms face fixed costs of selling in domestic and export markets, to explain phenomena such as the Balassa–Samuelson effect. Since only relatively more productive firms are profitable enough to engage in trade, they endogenously satisfy the precondition of the Balassa–Samuelson story that productivity gains are concentrated in the traded goods sector.

Loose Ends

International trade can influence macroeconomic activity through other channels. For example, as highlighted in endogenous growth models, technological progress may depend on incentives to undertake R&D and innovate, which, in turn, may depend on externalities or spillover effects from greater markets provided by international trade (Grossman and Helpman 1991). Greater openness to trade can also complicate the optimal conduct of monetary policy because of the impact of the exchange rate on real activity and inflation. Clarida et al. (2001) show how more openness to international trade can influence a central bank following an optimal policy feedback rule to raise the domestic interest rate more aggressively in response to inflation pressures. Lastly, trade may serve as a transmission channel through which financial crises may spread contagiously across countries (Glick and Rose 1999).

See Also

- ▶ Growth and International Trade
- ▶ International Real Business Cycles
- ▶ International Trade and Heterogeneous Firms
- ▶ International Trade Theory
- ▶ Marshall–Lerner Condition
- ▶ New Open Economy Macroeconomics
- ▶ Trade Costs
- ▶ Tradable and Non-tradable Commodities

Bibliography

Backus, D., P. Kehoe, and F. Kydland. 1992. International real business cycles. *Journal of Political Economy* 100: 745–775.

Baxter, M. 1995. International trade and business cycles. In *Handbook of international economics*, ed. G. Grossman and K. Rogoff, Vol. 3. Amsterdam: North-Holland.

Baxter, M., and M. Kouparitsas. 2005. Determinants of business cycle comovement: A robust analysis. *Journal of Monetary Economics* 52: 113–157.

Bergin, P., R. Glick, and A. Taylor. 2006. Productivity, tradability, and the long-run price puzzle. *Journal of Monetary Economics* 53: 2041–2066.

Betts, C., and M. Devereux. 1998. Exchange rate dynamics in a model of pricing to market. *Journal of International Economics* 50: 215–244.

Burstein, A., C. Kurz, and L. Tesar. 2005. International trade, production sharing and the transmission of business cycles. Working paper, University of California at Los Angeles.

Burstein, A., M. Eichenbaum, and S. Rebelo. 2006. The importance of nontradable goods' prices in cyclical real exchange rate fluctuations. *Japan and the World Economy* 18: 247–253.

Clarida, R., J. Gali, and M. Gertler. 2001. Optimal monetary policy in open versus closed economies: An integrated approach. *American Economic Review* 91: 248–252.

Corsetti, G., and P. Pesenti. 2001. Welfare and macroeconomic interdependence. *Quarterly Journal of Economics* 116: 421–445.
Dornbusch, R., S. Fischer, and P. Samuelson. 1977. Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods. *American Economic Review* 67: 823–839.
Frankel, J., and A. Rose. 1998. The endogeneity of the optimum currency area criteria. *Economic Journal* 108: 1009–1025.
Frankel, J.A., and A. Rose. 2002. An estimate of the effect of common currencies on trade and income. *Quarterly Journal of Economics* 117: 437–466.
Ghironi, F., and M. Melitz. 2005. International trade and macroeconomic dynamics with heterogenous firms. *Quarterly Journal of Economics* 120: 865–915.
Glick, R., and A. Rose. 1999. Contagion and trade: Why are currency crises regional? *Journal of International Money and Finance* 18: 603–617.
Grossman, G., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge: MIT Press.
Kose, M., and K.-M. Yi. 2006. Can the standard international business cycle model explain the relation between trade and comovement? *Journal of International Economics* 68: 267–295.
Obstfeld, M., and K. Rogoff. 1995. Exchange rate dynamics redux. *Journal of Political Economy* 103: 624–660.
Obstfeld, M., and K. Rogoff. 2000. The six major puzzles in international macroeconomics: Is there a common cause? In *NBER macroeconomics annual*, ed. B. Bernanke and K. Rogoff. Cambridge: MIT Press.
Stock, J., and M. Watson. 2005. Understanding changes in international business cycle dynamics. *Journal of the European Economic Association* 3: 968–1006.
Tille, C. 2001. The role of consumption substitutability in the international transmission of monetary shocks. *Journal of International Economics* 53: 421–444.
Warnock, F. 2003. Exchange rate dynamics and the welfare effects of monetary policy in a two-country model with home-product bias. *Journal of International Money and Finance* 22: 343–363.

Macroeconomic Forecasting

Mark W. Watson

Keywords

Bank of England; Dynamic adjustment; Dynamic stochastic general equilibrium models; Expectations; Federal Reserve System; Forecasting; Leading indicators index; Macroeconomic forecasting; Mean squared forecast error; Mitchell, W.C.; Probability theory; Structural econometric models; Time series analysis; Vector autoregressions

JEL Classifications

D4; D10

Macroeconomic forecasts are 'guesses' of the future values of important macroeconomic aggregates such as GDP, inflation, or the unemployment rate. These forecasts inform the decisions of business, policymakers, investors, and consumers. Macroeconomic forecasts are regularly constructed by government agencies and private companies. For example, every quarter the Bank of England publishes its *Inflation Report*, which contains forecasts of inflation over the next three years. Federal Reserve policymakers also rely on forecasts from the Green Book; however, unlike the Bank of England, the Fed does not release its forecasts to the public. The Federal Reserve Bank of Philadelphia summarizes private sector macroeconomic forecasts for the United States in its quarterly *Survey of Professional Forecasters*.

Macroeconomic forecasts are constructed using a variety of methods. These methods can be grouped into four categories: (1) leading indicator indexes; (2) structural econometric models; (3) time series models; and (4) judgement.

The origin of leading indicator indexes can be traced to the 1930s when, at the request of the US Secretary of the Treasury, Wesley Mitchell proposed a set of variables that historically had moved in anticipation of the business cycle. Averages of these leading indicators are an index of leading indicators. Such an index was constructed in the United States for several years by the Department of Commerce and is now maintained and published monthly by the Conference Board, which also publishes leading indicator indexes for several other countries.

Structural econometric models construct forecasts using dynamic relationships suggested by economic theory and estimated by statistical methods. Work on these models by Tinbergen, Klein and Haavelmo resulted in Nobel prizes for these researchers in 1969, 1980 and 1989

respectively. Large-scale structural models with hundreds of equations were developed in the 1960s and early 1970s, but forecast failures in the 1970s led researchers to question both the economic theory used in the models and the statistical procedures used to fit the models' equations. Refinements in theory (notably the importance of expectations and dynamic adjustment) and statistical methods (notably time series methods) are incorporated in the current generation of large-scale structural models. Currently, there is a significant research effort aimed at constructing small-scale structural models ('dynamic stochastic general equilibrium' models) for policy evaluation and forecasting.

Time series models use serial correlation (or persistence) in variables to construct forecasts. For example, a simple autoregressive model (AR) has the form $y_t = \alpha + \varphi y_{t-1} + \varepsilon_t$, where y_t is an economic variable of interest and ε_t is a zero-mean serially uncorrelated random shock. When φ is positive (negative), larger than average values of y_{t-1} tend to be associated with larger (smaller) than average values of y_t. Thus, an autoregressive forecast of y_{T+1} using data through time T is $y_{T+1|T} = \alpha + \varphi y_T$. Many macroeconomic variables display short-run dependence, and time series models typically produce more accurate short-run forecasts than other forecasting methods. Time series models have been developed to construct forecasts based on linear and nonlinear dependence properties in macroeconomic variables, and multivariate time series models, such as vector autoregressions (VARs), are widely used for short-horizon macroeconomic forecasting.

Professional forecasters also rely on judgement when constructing their forecasts. That is, while the macroeconomic forecasts published by the Bank of England or the Fed's Green Book forecasts rely on econometric models, the forecasts are not identical to model-based forecasts. Professional forecasters typically use judgement to adjust model-based forecasts. These adjustments – sometimes called 'add-factors' – allow forecasters (so they argue) to incorporate information that is not captured in the economic model. As an empirical matter, good judgement appears to improve the accuracy of model-based forecasts.

Much of the theory of forecasting can be derived from elementary concepts in probability theory. Let y_{T+1} denote the variable to be forecast and X_T denote a set of variables to be used for constructing the forecast. In general, X_T will include y_T, y_{T-1}, and longer lags, as well as current and lagged values of other series. Let $g(X_T)$ denote the forecast or 'guess' of y_{T+1} constructed from X_T, where good choices of $g(.)$ lead to more accurate forecasts. The forecast error is $e_{T+1} = y_{T+1} - g(X_T)$, and accuracy can be measured by mean squared forecast error (MSFE), where the conditional $\text{MSFE} = E\left(e_{T+1}^2 | X_T\right)$. A fundamental result from probability theory is that $E\left(e_{T+1}^2 | X_T\right)$ is minimized using $g(X_T) = E(y_{T+1}|X_T)$; that is, the regression (conditional expectation) produces the minimum mean squared forecast error.

A key implication of this theoretical result is that more information is always better – that is, it never hurts to include more variables in X_T, and the information in these additional variables will often reduce the MSFE. But, this result assumes that the regression function $E(y_{T+1}|X_T)$ is known, and in practice this function must be estimated using sample data. Including many variables in X_T means that many parameters must be estimated to characterize the regression function, and estimating a large number of parameters leads to statistical estimation error that increases the MSFE. This trade-off between including more variables in X_T to capture more information about y_{T+1} and the increased statistical error associated with estimating additional parameters for the forecasting model is one of the major practical problems in forecasting.

Another major problem is the temporal stability of the forecasting model. That is, the regression $E(y_{T+1}|X_T)$ might change over time, so that a regression estimated using past data might provide poor forecasts for future values of y_T. These two problems – developing methods for forecasting using many past variables and problems associated with instability – are active areas of current research. The relevant chapters in Elliott et al. (2006) summarize current research on these and other important topics in economic forecasting.

See Also

- ▶ Time Series Analysis
- ▶ Vector Autoregressions

Bibliography

Bank of England. 2007. *Inflation report*. Online. Available at http://www.bankofengland.co.uk/publications/inflationreport/index.htm. Accessed 19 Feb 2007.

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

Federal Reserve Bank of Philadelphia. 2007. *Survey of professional forecasters*. Online. Available at http://www.phil.frb.org/econ/spf. Accessed 19 Feb 2007.

Macroeconomics, Origins and History of

Robert W. Dimand

Abstract

Macroeconomics, the analysis of economic aggregates, became a recognized field with Keynes's *General Theory* (1936) and its mathematical and diagrammatic reformulations, and the macroeconometric modelling pioneered by Tinbergen and Frisch. Macroeconomics grew out of two long-standing traditions: business cycle analysis from Jevons and Juglar to Mitchell, and monetary theory, building on the work of Hume, Thornton, Ricardo, Wicksell, and Fisher, supplemented by the circular flow analysis of Quesnay and Marx.

Keywords

Austrian economics; Bank of England; Banking School, Currency School, Free Banking School; Bauer, O.; Bergmann, Eugen von; Beveridge, W. H.; Bimetallism; Bodin, J.; Burns, A. F.; Business cycles; Calculus of variations; Cantillon, R.; Capital accumulation; Central banking; Champernowne, D. G.; Circular flow; Cobb–Douglas functions; Consumption smoothing; Covered interest parity; Cowles Commission; Crowding out; Deflation; Depressions; Dunlop, J. T.; Endogenous growth; Equation of exchange; Excess demand and supply; Exploitation; Fel'dman, G. A.; Fisher ideal index; Fisher, I.; Fleming, J. M.; Friedman, M.; Frisch, R. A.; Fundamental uncertainty; German hyperinflation; Gervaise, I.; Gold standard; Growth theory; Haavelmo, T.; Hammarskjold, D.; Harrod, R. F.; Hawtrey, R. G.; Hayek, F. A.; Hicks, J. R.; Hoarding; Hume, D.; Income–expenditure analysis; Increasing returns; Index numbers; Index of leading indicators; Inflation; Innovations; Institute of World Economics (Germany); Insurable risk; International trade, theory of; Involuntary unemployment; IS–LM model; Jevons, W. S.; Jones, E. D.; Juglar, C.; Kaldor, N.; Kalecki, M.; Keynes, J. M.; Keynesian Revolution; Keynesianism; Kock, K.; Klein, L. R.; Kondratieff, N.; Koopmans, T. C.; Labour theory of value; Lange, O. R.; Laspeyres indexes; Law of markets; Lerner, A. P.; Lindahl, E. R.; Liquidity preference; Locke, J.; Lundberg, E. F.; Luxemburg, R.; Macroeconomics, origins and history of; Malthus, T. R.; Marshall, A.; Marx, K. H.; Meade, J. E.; Mercantilism; Mill, J.; Mill, J. S.; Mises, L. E. von; Mitchell, W. C.; Modigliani, F.; Monetarism; Money illusion; Mundell, R.; Myrdal, G.; National Bureau of Economic Research; National Institute of Economic and Social Research (UK); Natural rate and market rate of interest; Neoclassical; New classical macroeconomics; New Deal; New Keynesian macroeconomics; Ohlin, B. G.; Paasche index; Patinkin, D.; Periodogram; Phillips curve; Physiocracy; Pigou, A. C.; Post Keynesian economics; Price Revolution; Profit sharing; Public works; Purchasing power parity; Quantity theory of money; Quesnay, F.; Ramsey, F. P.; Recessions; Reddaway, B.; Representative agent; Ricardo, D.; Robbins, L. C.; Robertson, D. H.; Rules versus discretion; Samuelson, P. A.; Say, J.-B.; Say's Equality; Say's Identity; Say's Law; Simons, H. C.;

Simultaneous equations models; Slutsky, E.; Smith, A.; Specie-flow mechanism; Spending multiplier; Stabilization policy; Sunspots; Tarshis, L.; Taylor, F. I.; Thornton, H.; Time series analysis; Timlin, M.; Tinbergen, J.; Tobin, J.; Tooke, T.; Trade cycle theory; Uncovered interest parity; Underconsumptionism; Veblen, T.; Vector autoregressions; Velocity of circulation; Walras's Law; Warburton, C.; Wicksell, J. G. K.; Wilson, E. B.; Young. A. A

JEL Classifications
B2

Macroeconomics analyses a whole economy or economies, dealing with aggregate output and employment, the price level and interest rate, rather than with the prices or quantities of particular commodities. It became a recognized field as textbooks and course offerings responded to John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936; 1971–89, vol. 7), to the mathematical and diagrammatic reformulations of Keynes by David Champernowne, Brian Reddaway, Roy Harrod, J. R. Hicks, James Meade, Oskar Lange, Mabel Timlin and Franco Modigliani (Hicks 1937; Young 1987), and to the first aggregate econometric models such as Tinbergen (1939). Ragnar Frisch (1933) introduced the terms 'macrodynamics' and 'macroanalysis', and his distinction between macroanalysis and microanalysis is the same as the subsequent distinction between macroeconomics and microeconomics. Michal Kalecki (1935) first used 'macrodynamic' in a title, and by the time that Lawrence Klein (1946) used 'macroeconomics' in the title of a journal article, he presumed that its meaning would be clear to his readers. But just as Molière's *bourgeois gentilhomme* spoke prose long before he knew he was doing so, economists wrote macroeconomics long before they called it by that name. Macroeconomics grew out of two long-standing traditions within economics: business cycle analysis and the theory of money.

Macroeconomic Themes in Pre-classical and Classical Political Economy

The quantity theory of money is the oldest surviving theory in economics, yet remained, in David Laidler's (1991a) phrase, 'always and everywhere controversial' (primarily over whether changes in the quantity of money are exogenous or endogenous). Holding that a change in the money supply will ultimately change prices in the same proportion, the quantity theory was first used in the 16th century by Martin Navarro de Azpilcueta (writing in Latin as Navarrus) and other scholastics at the University of Salamanca (Grice–Hutchinson 1952), and then by Jean Bodin in France, to explain the 'Price Revolution', the inflation following the inflow of silver from the Spanish colonies in the New World. John Locke, Richard Cantillon and Isaac Gervaise contributed to understanding the velocity of circulation and the adjustment of international payments (Vickers 1959). The economic essays in David Hume's *Political Discourses* (see Hume 1752) mark a high-point of pre-classical monetary economics (see Humphrey 1993). Hume's analysis of the specie-flow mechanism of adjustment under the gold standard showed that an increase in the quantity of gold in one country would increase prices and spending in that country, causing a trade deficit and gold inflow until balance of payments of equilibrium was restored with the world's gold distributed among countries in proportion to their demand for real money balances. Hume's specie-flow mechanism provided a crushing rejoinder to mercantilist schemes for increasing the amount of gold in a country by promoting exports and restricting imports. Such tariffs, quotas and subsidies would distort resource allocation without producing a lasting trade surplus, and would raise prices rather than the real wealth of a nation. Hume recognized that an increased money supply would provide a temporary stimulus to real output, which would fade as prices and wages adjusted. While Hume linked each country's price level to that country's money stock and emphasized relative price effects on trade balances, his younger contemporary and friend

Adam Smith anticipated the monetary approach to the balance of payments by assuming purchasing power parity (with the world price level set by the world gold stock and world demand for real money balances) with adjustment taking place, not through relative price changes, but through the direct effect of a nation's excess demand for or supply of money on spending, hence on the balance of payments and on the country's stock of gold.

Keynes's *General Theory* revived interest in the debate in the years after the Napoleonic Wars about the possibility of a general glut of commodities. Keynes deplored the victory of David Ricardo's sharper analysis and endorsement of Say's (or James Mill's) Law of Markets over what Keynes regarded as Thomas Robert Malthus's deeper (but fuzzier) insight that insufficient effective demand could result in an excess supply of labour without an excess demand for any good (other than money). Malthus's insight was obscured by his failure to distinguish between a decision to save and a decision to invest, and hence to see the significance of hoarding. Statements of the Law of Markets by classical economists were more varied and complex, often subtler, and sometimes confused and contradictory than Keynes suggested in short quotations from the classics, which sometimes misled when taken out of context (see Link 1959; Corry 1962, on the macroeconomics of English classical economists and their critics, and Sowell 1972, on Say's Law). John Stuart Mill and others searched for a statement of the Law of Markets that would be the stronger truism that Oskar Lange later labelled as Say's Equality (if each and every commodity market is in equilibrium, then the sum of excess demand over all commodity markets much add to zero) but weaker than what Lange called Say's Identity, that excess demand for all commodity markets (that is, all markets except money) always sums to zero for any set of prices, regardless of whether any individual market is in equilibrium. Say's Identity, taken together with the adding up of budget constraints that Lange termed Walras's Law, implies that the money market always clears for any prices, leaving the absolute level of prices indeterminate. The policy implications that classical economists drew from their analysis are also more varied and pragmatic than the later textbook caricature: Jean-Baptiste Say recommended public works as a temporary response to unemployment during periods of adjustment, and criticized Ricardo for ignoring the possibility that savings might be hoarded if investment opportunities were inadequate. Ricardo, whose economic writings had begun with a pamphlet arguing that the premium on bullion demonstrated the wartime overissue and depreciation of inconvertible banknotes, was willing after the end of the war to support restoration of gold convertibility at the depreciated parity, rather than deflation to restore the pre-war parity. Henry Thornton (1802) introduced the concept of the central bank as the lender of last resort to support solvent but illiquid banks against bank runs. The proper role, if any, of the Bank of England generated prolonged controversy among the Banking, Currency, and Free Banking Schools in the first three quarters of the 19th century, producing analyses of lasting significance for monetary economics (Smith 1936; Fetter 1965).

François Quesnay's *Tableau Economique,* the crowning achievement of Physiocratic economics in France at the time of Hume and Smith, represented the circular flow of income and spending. It was not taken up by the mainstream of British and French classical political economy, but, a century after Quesnay, the *Tableau Economique* inspired Karl Marx's schemes of simple and expanded reproduction in the second volume of *Capital* (published posthumously in 1885), relating output and reinvestment rates in Department I (capital goods) and Department II (wage goods). For decades, this pioneering two-sector growth model was used only by Marxist economists such as Rosa Luxemburg and Otto Bauer constructing models of the supposed inevitable breakdown of capitalism, and then in 1928 by G. A. Fel'dman, proposing a growth theory for a planned economy. Fel'dman's articles were part of a false dawn of modern growth theory, appearing in the same year as the December 1928 issue of the *Economic Journal* that contained Allyn Young on increasing returns and

economic progress (inspired by Adam Smith) and Frank Ramsey's application of calculus of variations to optimal capital accumulation by a representative agent, but by 1930 Young and Ramsey were dead and Fel'dman had vanished in Stalin's purges (see Fel'dman, Ramsey and Young in Dimand 2002, vol. 3, and Bauer in vol. 5). Neoclassical hostility to Marx's theory of value and exploitation led to neglect of his contribution to growth theory, just as classical rejection of the Physiocratic doctrine of the exclusive net productivity of agriculture diverted attention from the circular flow. Marx also analysed the cyclical fluctuation of the profit rate around a downward trend, with cyclical troughs in the profit rate causing layoffs that force down wages by swelling the reserve army of the unemployed and cyclical peaks in the profit rate leading to realization crises as redistribution away from wages reduces demand for output (since Marx rejected Say's Law). However, his analysis of the increasing severity of crisis (as of the downward trend of the profit rate) was conducted within the special terminology and assumptions of his labour theory of value, which limited its influence on the mainstream of economics.

Business Cycles

Recognition of the more or less periodic recurrence of crises and prosperity goes back at least to Thomas Tooke's discussion in 1823 of 'waves' in prices (Arnon 1991), the beginning of the vast literature on cyclical fluctuations most conveniently sampled in the multi-volume anthologies of O'Brien (1997), Hagemann (2001), and Boianovsky (2005) and in the encyclopedia of Glasner (1997). Clément Juglar (1862) and W. Stanley Jevons (1884, collecting essays written from 1862 to 1882) advanced the analysis of economic fluctuations as periodic oscillations to a higher level, surpassing earlier descriptive and classificatory works (such as Max Wirth's *Geschichte der Handelskrisen* in 1858) and displacing the perception of crises as the result of occasional events. Jevons built upon Hyde Clarke's 1847 suggestion of a meteorological cause for the recurrence of crises every ten years or so (Hyde Clarke also perceived multiple, overlapping cycles, including a longer period of 54 years, anticipating Kondratiev). Jevons's sunspot theory of the trade cycle has so fallen out of favour that the term 'sunspots' in now used in the field of business cycles to refer to any intrinsically irrelevant variables (and even the term 'business cycles' is no longer taken to imply that fluctuations are in fact periodic cycles). This is unfair to Jevons, who was following the accepted meteorology of his era, which held that the cycle in solar activity affected weather. Cycles in weather would affect harvests, which, in a still largely agricultural world economy, would affect all economic sectors. Jevons's sunspot theory, together with his warnings about the impending exhaustion of coal, did much more than his marginal utility analysis of relative prices to persuaded the British Association for the Advancement of Science that economics was sufficiently scientific for Section F to remain in the Association. Nonetheless, as Wesley Mitchell (1927, p. 384) remarked,

> Jevons had an admirably candid mind; yet in 1875, when the sun-spot cycle was supposed to last 11.1 years, he was able to get from Thorold Rogers' *History of Agriculture and Prices in England* a period of 11 years in price fluctuations, and when the sun-spot cycle was revised to 10.45 years he was able to make the average interval between English crises 10.466 years.

Jevons was misled by the belief that an economic cycle must have a cause that is itself cyclical, but, as Knut Wicksell put it, the motion of a rocking horse does not resemble the motion of the stick that started it rocking (cited by Frisch 1933). Jevons's sunspot theory has distracted attention from such lasting contributions as the seasonal cycle (in his essay on the annual autumnal pressure on the Bank of England) and his use of index numbers to trace the effects of the Australian and California gold discoveries.

Wesley Mitchell (1913, 1927) was the leading figure in the statistical approach to business cycle analysis. In 1920, Mitchell founded the National Bureau of Economic Research (NBER), which was the model for institutes of business cycle or conjuncture research in Berlin, Vienna (directed

first by Friedrich Hayek and then by Oskar Morgenstern), Belgium, Sofia, Moscow (directed by Nikolai D. Kondratiev, the theorist of long waves), and Warsaw (where Kalecki worked), Britain's National Institute of Economic and Social Research, and the Institute of World Economics in Kiel. Although his Columbia lectures on types of economic theory were famous, Mitchell was sceptical about taking any single explicit economic theory, such as the quantity theory of money or utility maximization, as a starting point, as he felt that many of the theories surveyed in Mitchell (1927) captured something of the truth, but none the whole truth. Mitchell was influenced by his teacher at the University of Chicago, the institutionalist Thorstein Veblen (1904), who coined the term 'neoclassical' to describe the sort of Marshallian economics of which he disapproved. Mitchell and Arthur F. Burns (his successor directing the NBER) concentrated on investigating the statistical properties of time series, looking for patterns of leads and lags and for superimposed cycles of different periods and amplitudes. The widely reported index of leading indicators continues the original NBER approach.

Sir William Beveridge, director of the London School of Economics, used the periodogram, an early version of spectral analysis, to decompose wheat prices into 19 cycles with periods varying from 2.735 years to 68 years (Beveridge 1921, 1922). Finding so many cycles led sceptics, such as Harvard statistician E.B. Wilson, to wonder whether there were any truly periodic oscillations in economic time series (apart from seasonality), since with enough cycles any series could be represented as a summation of cycles. Eugen Slutsky (1937, originally published in Russian in 1927), used a moving average of the last three digits of the winning Moscow lottery numbers to show that summation of random series could produce apparent cycles. Slutsky (1937) and Frisch (1933) influenced economists to consider fluctuations as oscillatory responses to random shocks (real or monetary), turning away from the emphasis of Jevons, Juglar, Mitchell, Beveridge and Kondratiev on underlying cycles. Cowles Commission director Tjalling Koopmans (1947) denounced Burns and Mitchell's *Measuring Business Cycles* (1946) as 'Measurement without Theory', and argued instead for simultaneous equation macroeconometric models, with the equations identified by exclusionary restrictions derived from a priori economic theory. Koopmans's Chicago colleague Milton Friedman (whose Columbia dissertation had been supervised by Burns) responded by writing down a formal model representing Mitchell's business cycle analysis (Friedman 1952, Section III and Appendix, pp. 257–82). The vector autoregressions (VAR) of Christopher Sims (1980) marks a return (with more modern statistical techniques) to the NBER approach of investigating the statistical properties of macroeconomic time series with only limited reliance on a priori restrictions drawn from theory.

1886 and All That: The Dawn of Modern Monetary Macroeconomics, 1886–1914

Around 1886, during a period of depression, analysis of cycles and crises acquired a new emphasis on fluctuation of employment as the problem and variations in the general price level as a preventable cause. Carroll Wright (1886) devoted his first annual report as US Commissioner of Labor to a statistical study, *Industrial Depressions*, finding such depressions to be largely contemporaneous across manufacturing nations and advocating profit-sharing to mitigate the severity of fluctuations (a proposal independently rediscovered nearly a century later by Martin Weitzman). In the same year, Britain had a Royal Commission on the Depression of Trade and Industry, chaired by Lord Iddesleigh (Stafford Northcote) and including Professor Bonamy Price of Oxford but most notable for the evidence of Professor Alfred Marshall of Cambridge. In his evidence to that inquiry and to the Gold and Silver Commission of 1887–8 (both reprinted in Marshall 1926, edited by Keynes), and in a paper to the Industrial Remuneration Conference of 1885, Marshall considered how far remediable causes adversely affect continuity of employment. This led him to suggest 'Remedies for Fluctuations of General Prices' in the *Contemporary Review* in March 1887 (reprinted in Pigou

1925), revising Ricardo's ingot plan to make the monetary unit a claim on a fixed weight of gold plus a fixed weight of silver, a step toward pegging the monetary unit to a basket of commodities (Irving Fisher's compensated dollar). This symmetallism proved incomprehensible to bimetallists, who persisted in plans that would require pegging the relative price of gold and silver. Also in 1886 (two years after writing the introduction to the posthumous collections of Jevons's *Investigations in Currency and Finance),* Herbert Foxwell published a lecture on *Irregularity of Employment and Fluctuations of Prices* (in Dimand 2002, vol. 1). Like his colleague Marshall (both were fellows of St John's College, Cambridge), Foxwell emphasized fluctuations in employment as the crucial challenge posed by economic instability, and argued that the problem 'How to secure greater industrial stability' could be reformulated as 'How to diminish price fluctuations'. A young Swedish student named Knut Wicksell attended Foxwell's lectures at University College London in 1886. As Wesley Mitchell (1927, p. 7) observed, 'Before the end of the nineteenth century there had accumulated a body of observations and speculations sufficient to justify the writing of histories of the theories of crises': Eugen von Bergmann's *Die Wirtschaftskrisen: Geschichte der nationalökonomischen Krisentheorien,* published in Stuttgart in 1895, and E. D. Jones's *Economic Crises,* published in New York in 1900 (see also Barnett 1941). In 1909, the year of Beatrice and Sidney Webb's Minority Report of the Poor Law Commission and of the first edition of Beveridge's *Unemployment,* the London School of Economics published a 71-page bibliography of unemployment and the unemployed by F. Isabel Taylor.

The cover of David Laidler's *The Golden Age of the Quantity Theory* (1991b) shows the three economists who dominated monetary economics before the First World War, making the case for monetary shocks and imbalances as the avoidable source of fluctuations: Alfred Marshall, Knut Wicksell and Irving Fisher.

In *Interest and Prices* in 1898 and then in his *Lectures on Political Economy* (1915), Wicksell distinguished the market rate of interest, set by the banking system, from the natural rate, the interest rate at which desired saving and investment would balance and the price level would not change. If a technical innovation raises the natural rate, or the banking system lowers the market rate, it will be profitable for entrepreneurs to borrow for new investment projects as long as the natural rate exceeds the market rate, causing (in a pure credit economy with no cash drain) a cumulative inflation. If the market rate exceeded the natural rate, a cumulative deflation would ensue. Although he considered himself a quantity theorist following in the footsteps of Ricardo, Wicksell was a pioneer in analysing a pure credit economy, not anchored by gold or other base money, which is why Michael Woodford deliberately chose Wicksell's title *Interest and Prices* for his 2003 treatise analysing a world in which financial innovation has greatly reduced the role of cash and bank reserves. Wicksell's economic contributions (which included using what came to be called the Cobb–Douglas production function four years before Cobb and Douglas) were continued by a Stockholm School including Dag Hammarskjold, Karin Kock (1929), Erik Lindahl (1939), Erik Lundberg (1937), Gunnar Myrdal (1939) and Bertil Ohlin – a list including three Nobel laureates (two in economics, one in peace) and four Swedish cabinet ministers. The Stockholm economists later expressed confidence that, even if Keynes had never written *The General Theory*, they would have discovered it themselves, but Don Patinkin (1982) expressed doubt, because the focus of Wicksell's heirs was on price dynamics, not the equilibrium level of employment and national income. Keynes's earlier *Treatise on Money* in 1930 (1971–89, vols. 5 and 6) was much more Wicksellian than *The General Theory* in its emphasis on cumulative inflation or deflation when the interest rate does not equate planned investment to planned saving.

J. Bradford De Long (2000) writes:

> The story of 20th century macroeconomics begins with Irving Fisher. In his books *Appreciations and Interest* (1896), *The Rate of Interest* (1907), and *The Purchasing Power of Money* (1911) [Fisher 1997, vols. 1, 3, and 4], Fisher fueled the intellectual fire that much later became monetarism. To understand the determination of prices and interest rates and the course of the business cycle, monetarism holds,

M

> look first (and often last) at the stock of money – at the quantities in the economy of those assets that constitute readily spendable purchasing power. . . . It is true that the ideas that we see as necessarily producing the quantity theory of money go back to David Hume, if not before. But the equation-of-exchange and the transformation of the quantity theory of money into a tool for making quantitative analyses and predictions of the price level, inflation, and interest rates was the creation of Irving Fisher.

In *Appreciation and Interest,* Fisher argued that the difference between interest rates expressed in two standards (money and commodities, gold and silver, dollars and francs) is the expected rate of appreciation of one standard in terms of the other, deriving from this uncovered interest parity between two countries, the expectations theory of the term structure of interest rates, and the Fisher relation that nominal interest is the sum of real interest and expected inflation (plus a cross-product term). In *The Rate of Interest,* Fisher introduced the Fisher diagram, showing the optimal smoothing of consumption over two periods (assuming perfect credit markets) and an individual's saving or dissaving in each period. In *The Purchasing Power of Money*, Fisher (with his former student Harry G. Brown) upheld the quantity theory both against bimetallists who predicted permanent real benefits from expanding the money supply and against hard-money opponents of bimetallism (notably J. Laurence Laughlin of the University of Chicago), who denied the path of US prices could be explained by changes in the money supply. Fisher and Brown explained economic fluctuations by the slow adjustment of nominal interest to monetary shocks during 'transition periods' (lasting perhaps ten years), so that fluctuations could be avoided either by educating the public against what Fisher later termed 'the money illusion' (so that expected inflation and hence nominal interest would adjust to monetary shocks, leaving real interest unaltered) or by a monetary policy rule of varying the exchange rate (the dollar price of gold) to hold constant a price index (for which Fisher later proposed the Fisher ideal index, the geometric mean of the Paasche and Laspeyres indexes). Fisher's 1926 article, 'A statistical relation between unemployment and price changes' (in Fisher 1997, vol. 8), correlated unemployment with a distributed lag of past price level changes (as a proxy for expected inflation), and was reprinted in the *Journal of Political Economy* in 1973 under the heading 'Lost and Found: I Discovered the Phillips Curve – Irving Fisher'. Unlike Marshall in Cambridge and Wicksell in Stockholm, Fisher did not attract a school of disciples at Yale. Through his role in establishing the Econometric Society and the Cowles Commission, Fisher advanced his preferred economic methodology of formal theorizing using mathematical and statistical techniques, but his contributions to monetary economics and economic fluctuations (like those of Hayek, Hawtrey, and many others) were long overshadowed by Keynes's *General Theory,* notwithstanding Keynes's acknowledgement of Fisher as his intellectual great-grandparent in appreciating the real effects of monetary changes.

Although Alfred Marshall's *Money, Credit and Commerce* was not published until 1923, the year before his death, parts of it were drafted as early as the 1870s, and his ideas had long circulated through his lectures, his evidence to official inquiries (gathered by Keynes in Marshall 1926), and the 'Cambridge oral tradition' of monetary theory (Eshag 1963; Bridel 1987; Laidler 1999). Marshall, his professorial successor A. C. Pigou, Pigou's successor D.H. Robertson (1926), the young J.M. Keynes, and Cambridge economics lecturers Frederick Lavington and J.R. Bellerby used a cash balance version of the quantity theory, relating the number of units of purchasing power the public wished to hold as cash to the level of income (in contrast to Fisher's logically equivalent version, which expressed the quantity theory in terms of the velocity of circulation of money).

Departure from the gold standard during the First World War and the post-war central European hyperinflations provided the occasion for the highest achievement of Marshallian monetary economics, Keynes's *Tract on Monetary Reform* in 1923 (Keynes 1971–89, vol. 4), an innovative work but one that innovated within the tradition established by Marshall. Keynes analysed inflation as a form of taxation of real money balances, identified as a social cost the consequent reduction in desired holdings of real money balances (M/P),

and introduced covered interest parity (the spread between forward and spot exchange rates equals the difference between interest rates in the two currencies). Keynes calculated that real money balances had fallen by 92 per cent during the German hyperinflation, as a result of the soaring opportunity cost of holding money. Others had mistakenly argued that since the price level (P) was rising faster than the money supply (M), monetary expansion could not be the cause of the price inflation, and the Reichsbank president Rudolf Havenstein promised that, with 38 new high-speed printing presses, the Reichsbank would be able to print enough money to catch up with the prices. Robertson (1926), then collaborating closely with Keynes, examined forced saving ('induced lacking' in Robertson's terminology) caused by inflation. Turning from inflation to deflation, Keynes wrote *The Economic Consequences of Mr. Churchill* (in Keynes 1971–89, vol. 9) to oppose Britain's return to the gold standard at the pre-war parity in 1925, arguing that restoration of the pre-war parity would require a reduction of prices and money wages that could be achieved only through prolonged unemployment (see June Flanders 1989, on the development of international monetary economics).

Keynesian Revolution and Monetarist Counter-Revolution

The Great Depression of the 1930s helped provide a receptive audience for John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936; 1971–89, vol. 7), which argued that involuntary unemployment could persist unless the government intervened with appropriate management of aggregate demand (Clarke 1988; Dimand 1988; Backhouse 1995). *The General Theory* challenged Lionel Robbins and Friedrich Hayek of the London School of Economics, who argued against expansionary fiscal and monetary policy and for letting the depression take its course, and William Beveridge, who held (until his conversion to Keynesianism) that the existing level of British unemployment could be fully accounted for by structural, frictional, and seasonal unemployment without invoking any deficiency of aggregate demand. To the rising generation of new economists, from Harvard students Paul Samuelson and James Tobin to LSE economists Abba Lerner and Nicholas Kaldor, Keynes offered a message of hope that depressions were curable and preventable without adopting a Soviet-style centrally planned economy. Attempts to dismiss or ignore Keynes (Burns and Mitchell 1946, mentioned Keynes in one sentence, in a footnote) were futile. Keynes provided an agenda for economists providing public policy advice and a framework for empirical, policy-oriented modelling, at a time when depression and war greatly expanded the role of governments.

Keynes's success in winning over the next generation of economists obscured the extent to which his contemporaries in economics shared his policy views rather than those of Robbins and Hayek: although Keynes used Pigou (1933) as the target of his attack on classical theory, he recognized how close they stood on practical policy. Even Ralph Hawtrey, the Treasury economist associated with the 'Treasury view' about crowding out and the ineffectiveness of fiscal policy, was convinced of the effectiveness of (and need for) stabilizing monetary policy (and contributed intriguing numerical examples to the development of the Kahn–Keynes spending multiplier; see Hawtrey 1932). Keynes's caricature in *The General Theory* of 'classical economists' from Ricardo to Pigou as upholders of a rigid version of Say's Law, denying any role to aggregate demand in explaining unemployment (in contrast to the superior insight but fuzzier analytics of mercantilists, Malthus, and the underconsumptionists Hobson and Mummery), was more widely noted than his subsequent clarification that he did not consider Fisher or Hawtrey or Robertson or Wicksell as classical. However, support for expansionary fiscal or monetary policy during the Depression did not necessarily imply anticipation of Keynesian economics: proposals circulated for emergency public works financed by cutting other government spending and for domestic monetary expansion while keeping the exchange rate fixed, and in the United

States the New Deal's National Recovery Administration was an attempt to raise price toward pre-Depression levels by restricting supply, rather than by stimulating demand. Keynes provided a framework within which the implications of such policies could be analysed. Independently of Keynes, starting from Marx and Rosa Luxemburg, Michal Kalecki in Poland developed a theory very close to Keynes's income–expenditure analysis, and in 1934 published in Polish a three-equation model of goods market equilibrium, money market equilibrium and aggregate supply. Patinkin (1982) argued that Kalecki was concerned with the dynamics of cyclical fluctuations, Keynes with determining the equilibrium level of income that equates saving to desired investment, and that Kalecki's 1934 essay (which Kalecki did not choose to be translated among his selected articles in 1966 and 1971, or refer to in other works) was not part of his central message.

The analytical framework that dominated macroeconomics for at least a quarter century after the Second World War was based on Keynes's aggregate supply and aggregate demand functions (generally with more attention to aggregate demand than to aggregate supply) and the small system of simultaneous equations behind the Hicks–Hansen IS/LM diagram, which included Keynes's money demand function (liquidity preference) and later substitutes for his consumption function (De Vroey and Hoover 2004). The system of equations representing Keynes's message in a form equivalent to IS/LM was a four-equation model in Keynes's Cambridge lectures in December 1933, attended by David Champernowne and Brian Reddaway, the first economists to use such a model in print, but Keynes did not include it in *The General Theory,* perhaps following Marshall's advice to use mathematics as a tool of inquiry but to then translate the analysis into English and burn the mathematics (Rymes 1989; Dimand 1988). The resulting framework (extended to open economies by Robert Mundell and J. Marcus Fleming in the 1960s) did not capture all of Keynes's message (or messages), notably his distinction between fundamental uncertainty and insurable risk. Econometric estimation of macroeconomic models was pioneered, independently of Keynes, by Ragnar Frisch, Jan Tinbergen and Trygve Haavelmo (and Keynes's review of the first volume of Tinbergen 1939, expressed severe scepticism), but it was taken up with enthusiasm by such Keynesians as Lawrence Klein. The claim in Chapter 2 of *The General Theory* that real and money wages move in opposite directions over the course of the cycle (and by implication, that real wages vary countercyclically) was challenged empirically by John Dunlop and Lorie Tarshis and on theoretical grounds by Michal Kalecki, leading Keynes in 1939 to acknowledge the cyclical pattern of real wages as an open question, which it remains to this day.

Milton Friedman and his students (Friedman 1956) offered a renewed quantity theory of money as a challenge to Keynesianism, claiming to follow a Chicago oral tradition of monetary theory. Certainly it drew on such Chicago landmarks as Henry Simons's 1936 argument for rules rather than authorities in monetary policy (reprinted in Simons 1948), but the intellectual inheritance from non-Chicago quantity theorists such as Irving Fisher and Clark Warburton (and even the young Keynes of *A Tract on Monetary Reform)* gradually came to be recognized. As Patinkin (1981) noted, a key element of Friedman's approach, the demand for money as a function of a small number of variables, originated in Keynes's *General Theory.* Although others had come close (in 1930, Fisher stated the marginal opportunity cost of holding cash balances), Keynes was the first to write the demand for money as a function of income and the interest rate. A further irony was that, though the spread of Keynesianism stemmed largely from its apparent ability to explain the Great Depression, the monetary interpretation of the Great Depression by Friedman and Schwartz (1963), as the consequence of mistaken Federal Reserve policy that permitted the US money supply to contract by a third, was crucial in persuading many economists of the explanatory power of monetarism, the revived form of the quantity theory. For an overview of the development of macroeconomics from Keynes through Friedman to the New

Classical and New Keynesian research programs (and the non-mainstream Post Keynesian and Austrian schools, from Keynesian fundamental uncertainty and Mises–Hayek trade cycle theory, respectively), enlivened with interviews with leading participants (see Snowdon and Vane 2005).

Recurring Themes

Certain issues reappear throughout the history of economics. Do fluctuations result from monetary disturbances, as Hawtrey (1913, 1932) and Fisher argued, or from real productivity shocks such as Schumpeterian innovations? Is unemployment best analysed as the functioning or malfunctioning of the labour market (as in Beveridge 1930; Hutt 1939) or in terms of the demand for and supply of output as a whole (Keynes)? Is there a role for demand management to offset instability caused by volatile private investment reflecting the fundamental uncertainty of future profitability (Keynes) or is government itself the source of instability (von Mises, Hayek)? Should a central bank follow a rule rather than having discretion (as Henry Simons asked in 1936), or need there even be a central bank (as Hayek's student Vera Smith asked the same year)? Are recessions undesirable and preventable disequilibrium phenomena, or, as Arthur Ellis (1879) and Friedrich Hayek (1931) held, are they a normal and necessary part of the equilibrium path of the economy? Should analysis of economic fluctuations should be primarily a study of the statistical properties of the fluctuations, as in Burns and Mitchell (1946) and decades later Sims's vector autoregressions, or should the analysis be explicitly grounded in formal economic theory? As macroeconomists continue to theorize, measure, test, and argue about these issues, they stand, knowingly or not, on the 'shoulders of giants' who discussed these questions before. De Long (2000, p. 83) notes that 'The New Classical research program walks in the footprints of Joseph Schumpeter's *Business Cycles* (1939) [and of Schumpeter 1912; Robertson 1915], holding that the key to the business cycle is the stochastic nature of economic growth [so that] the "cycle" should be analysed with the same models used to understand the "trend"' ,while the name of the New Keynesian research program (which emphasizes frictions that prevent instantaneous adjustment to nominal shocks) indicates its historical antecedents (although, as De Long points out, it also incorporates important features of Milton Friedman's contributions, such as emphasis on policy rules and on monetary rather than fiscal policy). Insights have sometimes long preceded the ability to formalize them; even Adam Smith's famous increasing returns through the division of labour, revived in Allyn Young's 1928 essay on economic progress, did not make its mark on the theories of international trade and endogenous growth until the last decades of the 20th century, when ways were devised to incorporate increasing returns to scale in formal models. The field has experienced major changes, as when Keynes made determining the equilibrium level of national income the central issue, or when monetarism posed inflation as the central problem instead of unemployment, or when attention shifted from fluctuations to long-term growth, but in each case the change was a transformation of a rich heritage.

Bibliography

Arnon, A. 1991. *Thomas Tooke, pioneer of monetary theory*. Aldershot: Edward Elgar.

Backhouse, R.E. 1995. *Interpreting macroeconomics: Explorations in the history of macroeconomic thought*. London/New York: Routledge.

Barnett, P. 1941. *Business-cycle theory in the United States, 1860–1900*. Chicago: University of Chicago Press.

Beveridge, W.H. 1921. Weather and harvest cycles. *Economic Journal* 21: 429–452.

Beveridge, W.H. 1922. Wheat prices and rainfall in Western Europe. *Journal of the Royal Statistical Society* 85: 412–459.

Beveridge, W.H. 1930. *Unemployment, a problem of industry (1909 and 1930)*. London: Longmans, Green.

Boianovsky, M. (ed.). 2005. *Business cycle theories: Selected texts 1860–1939,* vols. 5–8, London: Pickering & Chatto.

Bridel, P. 1987. *Cambridge monetary thought: The development of saving–investment analysis from Marshall to Keynes*. Basingstoke: Macmillan.

Burns, A.F., and W.C. Mitchell. 1946. *Measuring business cycles*. New York: NBER.
Clarke, H. 1847. *Physical economy: A preliminary inquiry into the physical laws governing the periods of famines and panics.* London (Cat. No. 34987.8, Kress Library, Harvard University).
Clarke, P. 1988. *The Keynesian revolution in the making 1924–1936*. Oxford: Oxford University Press.
Corry, B. 1962. *Money, saving and investment in English economics 1800–1850*. London: Macmillan.
De Long, J.B. 2000. The triumph of monetarism? *Journal of Economic Perspectives* 14(1): 83–94.
De Vroey, M., and K.D. Hoover (ed.). 2004. *The IS–LM Model: Its rise, fall, and strange persistence.* Durham: Duke University Press. *Annual supplement to history of political economy 36.*
Dimand, R.W. 1988. *The origins of the Keynesian revolution*. Aldershot/Stanford: Edward Elgar/Stanford University Press.
Dimand, R.W. (ed.) 2002. *The origins of Macroeconomics,* 10 vols. London/New York: Routledge.
Ellis, A. 1879. *The rationale of market fluctuations*. 4th ed. London: Effingham Wilson.
Eshag, E. 1963. *From Marshall to Keynes: An essay on the monetary theory of the Cambridge school*. Oxford: Basil Blackwell.
Fetter, F.W. 1965. *The development of British monetary orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.
Fisher, I. 1997. *The works of irving fisher*, 14 vols, ed. W. J. Barber assisted by R. Dimand and K. Foster, consulting ed. J. Tobin. London: Pickering & Chatto.
Flanders, M.J. 1989. *International monetary economics 1870–1960*. Cambridge: Cambridge University Press.
Friedman, M. 1952. The economist theorist. In *Wesley Clair Mitchell: The economic scientist*, ed. A.F. Burns. New York: NBER.
Friedman, M. (ed.). 1956. *Studies in the quantity theory of money.* Chicago: University of Chicago Press.
Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press for NBER.
Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honour of Gustav Cassell*. London: George Allen & Unwin.
Glasner, D. (ed.). 1997. *Business cycles and depressions: An encyclopedia.* New York: Garland.
Grice-Hutchinson, M. 1952. *The school of Salamanca: Readings in Spanish monetary theory 1544–1605*. Oxford: Clarendon Press.
Hagemann, H. (ed.). 2001. *Business cycle theories: Selected texts 1860–1939,* 4 vols. London: Pickering & Chatto.
Hawtrey, R.G. 1913. *Good and bad trade.* London: Constable. Reprinted, New York: A. M. Kelley, 1970 (with 1962 preface by author).
Hawtrey, R.G. 1932. *The art of central banking*. London: Longmans, Green.
Hayek, F.A. 1931. *Prices and production*. London: Routledge.
Hicks, J.R. 1937. Mr. Keynes and the classics: A suggested interpretation. *Econometrica* 5: 147–159.
Hume, D. 1752. In *Writings on economics*, ed. E. Rotwein, 1955. Madison: University of Wisconsin Press.
Humphrey, T.M. 1993. *Money, banking, and inflation: Essays in the history of monetary thought*. Aldershot// Brookfield: Edward Elgar.
Hutt, W.H. 1939. *The theory of idle resources*. London: Cape.
Jevons, W.S. 1884. Investigations in currency and finance, ed. H.S. Foxwell. London: Macmillan.
Juglar, C. 1862. *Des crises commerciales et de leur retour périodique en France, en Angleterre et aux Etats-Unis*. Paris: Guillaumin.
Kalecki, M. 1934. Trzy uktady [Three economic models]. *Ekonomista* 34: 54–70. Trans. C.A. Kisel. In M. Kalecki, Collected works of Michal Kalecki, volume 1: Capitalism, *Business cycles and full employment,* ed. J. Ostiatynski. Oxford: Clarendon Press, 1990.
Kalecki, M. 1935. A macrodynamic theory of the business cycle. *Econometrica* 3: 327–344.
Keynes, J.M. 1971–89. *Collected writings of John Maynard Keynes,* 30 vols., ed. D.E. Moggridge and E.A.G. Robinson. London/New York: Macmillan/ Cambridge University Press, for the Royal Economic Society.
Klein, L.R. 1946. Macroeconomics and the theory of rational behavior. *Econometrica* 14(2): 93–108.
Kock, K. 1929. *A study of interest rates.* Stockholm economic studies no. 1. London: P.S. King.
Koopmans, T.C. 1947. Measurement without theory. *Review of Economic Statistics* 29: 161–172.
Laidler, D. 1991a. *The golden age of the quantity theory.* Princeton: Princeton University Press.
Laidler, D. 1991b. The quantity theory is always and everywhere controversial – Why? *Economic Record* 67: 289–306.
Laidler, D. 1999. *Fabricating the Keynesian revolution: Studies of the inter-war literature on money, the cycle, and unemployment*. Cambridge: Cambridge University Press.
Laidler, D. 2003. Macroeconomics in retrospect: Selected essays. Cheltenham/Northampton: Edward Elgar.
Lindahl, E. 1939. *Studies in the theory of money and capital*. London: George Allen & Unwin.
Link, R. 1959. *English theories of economic fluctuations 1815–1848*. New York: Columbia University Press.
Lundberg, E. 1937. *Studies in the theory of economic expansion*. London: P.S. King.
Marshall, A. 1923. *Money, credit and commerce*. London: Macmillan.
Marshall, A. 1926. In *Official papers*, ed. J.M. Keynes. London: Macmillan.
Mitchell, W.C. 1913. *Business cycles*. Berkeley: University of California Press.
Mitchell, W.C. 1927. *Business cycles: The problem and its setting*. New York: NBER.

Myrdal, G. 1939. *Monetary equilibrium*. Trans. R. Bryce and N. Stolper. London: W. Hodge.
O'Brien, D.P. (ed.). 1997. *Foundations of business cycle theory,* 3 vols. Cheltenham/Brookfield: Edward Elgar.
Patinkin, D. 1981. *Essays on and in the Chicago tradition*. Durham: Duke University Press.
Patinkin, D. 1982. *Anticipations of the general theory? And other essays on Keynes*. Chicago: University of Chicago Press.
Pigou, A.C. (ed.). 1925. *Memorials of Alfred Marshall.* London: Macmillan.
Pigou, A.C. 1933. *Theory of unemployment*. London: Macmillan.
Robertson, D.H. 1915. *A study of industrial fluctuation*. London: P.S. King.
Robertson, D.H. 1926. *Banking policy and the price level*. London: P.S. King.
Rymes, T.K. 1989. *Keynes's lectures, 1932–35: Notes of a representative student*. Basingstoke/Ann Arbor: Macmillan/University of Michigan Press.
Schumpeter, J.A. 1912. *The theory of economic development.* Trans. R. Opie. Cambridge, MA: Harvard University Press, 1934.
Schumpeter, J.A. 1939. *Business cycles,* 2 vols. New York: McGraw-Hill.
Simons, H. 1948. *Economic policy for a free society.* Chicago: University of Chicago Press.
Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.
Slutsky, E. 1937. The summation of random causes as the source of cyclic processes. *Econometrica* 5: 105–146.
Smith, V.C. 1936. *The rationale of central banking*. London: P.S. King.
Snowdon, B., and H.R. Vane. 2005. *Modern macroeconomics: Its origins, development and current state*. Cheltenham: Edward Elgar.
Sowell, T. 1972. *Say's law: A historical analysis*. Princeton: Princeton University Press.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of great Britain,* ed. F.A. Hayek. New York*:* A.M. Kelley, 1965.
Tinbergen, J. 1939. *Statistical testing of business cycle Theories,* 2 vols. Geneva: League of Nations. Reprinted, New York: Agathon Press, 1968.
Veblen, T. 1904. *The theory of business enterprise*. New York: Charles Scribner's Sons.
Vickers, D. 1959. *Studies in the theory of money 1690–1776*. New York: Chilton.
von Mises, L. 1935. *The theory of money and credit.* Trans. H. Batson. London: Cape.
Wicksell, K. 1898. *Interest and prices.* Trans. R. Kahn. London: Macmillan, 1936.
Wicksell, K. 1915. *Lectures on political economy*, vol. 2. Trans. E. Claasen. London: Routledge, 1935.
Wirth, M. 1858. *Geschichte der Handelskrisen.* 4th edn, Frankfurt am Main, 1890.
Woodford, M. 2003. *Interest and prices: foundations of a theory of monetary policy.* Princeton: Princeton University Press.
Wright, C.D. 1886. *Industrial depressions: The first annual report of the commissioner of Labor,* March, 1886. Washington, DC: Government Printing Office. Reprinted, New York: Augustus M. Kelley, 1968.
Young, A.A. 1928. Increasing returns and economic progress. *Economic Journal* 38: 527–542.
Young, W. 1987. *Interpreting Mr. Keynes: The IS–LM enigma*. Oxford/Boulder: Polity Press/Westview.

Macroeconomics: Relations with Microeconomics

Peter Howitt

The lack of clear connection between macroeconomics and microeconomics has long been a source of discontent among economists. Arrow (1967) called it a 'major scandal' that neoclassical price theory cannot account for such macroeconomic phenomena as unemployment. Lucas and Sargent (1979) argued that Keynesian macroeconomics is 'fundamentally flawed' by its lack of a firm microfoundation. Countless students and practitioners alike have complained of the schizophrenic nature of a discipline whose two major branches project such radically different views of the world.

It is not hard to see why this lack of unity should bother economists. Fragmentary explanations are intellectually unsatisfying in any discipline, and are rightly labelled *ad hoc*. Theories that must be altered when moving from one application to another do not provide general covering laws and are liable to break down when new applications are tried or when new data arise.

The urge to close the micro–macro gap has been particularly strong among macroeconomic theorists, whose general desire for unity has been reinforced by at least three special factors. First, there is the reductionist methodological predisposition that economists of almost all persuasions share to some degree, according to which no explanation of economic phenomena is truly satisfactory if it does not reduce the phenomena to a question of individual actions by basic

decision-making units. Second, the lack of decisive empirical tests or experiments in economics has precluded demonstrations that macroeconomic theory is valid within a well-delineated domain of applicability, despite the difficulty of reconciling it with micro principles. This same factor has also forced economists to rely to a large extent upon introspection as a criterion and source of new ideas, a reliance which enhances their reductionist tendencies, since introspection is easier to apply if a theory is cast in terms of individual actions rather than in terms of broad social forces or primitive relationships among aggregate variables. Third, microeconomics was codified and given a well-articulated mathematical structure when macroeconomics was just emerging from its pre-analytical stage. Historically, one of the most fruitful strategies for macro theorists has been to borrow and apply the principles, conventions and techniques that have succeeded in the theoretically more advanced branch of the discipline.

Thus the quest for microfoundations has been a mainspring of development in macro theory. However, this does not mean that macro has been developing into a branch of applied micro. The forces tending to make macro theory conform more closely with micro principles have been opposed by equally important forces requiring those principles to be modified radically before being applied to macro questions.

More specifically, what has restrained the urge to apply micro principles is a widespread recognition that some of the most important macroeconomic phenomena manifest defects in the economic system that standard micro theory rules out with its basic assumption of equilibrium. In a state of general equilibrium, as traditionally conceived by micro theorists, all trading plans are costlessly coordinated by 'the market', whose operation is often heuristically personified by the Walrasian auctioneer. The auctioneer establishes prices such that plans are collectively compatible, with demand and supply equated for each commodity. He also ensures that, given this compatibility condition, all trading plans can be executed at no cost.

With the auctioneer at work, one individual's decisions interfere with another's only to the extent that they affect the vector of equilibrium prices. Thus the only constraint that social interactions are assumed to impose on the formation of trading plans is the single budget constraint requiring the value of purchases not to exceed the value of sales. No one need concern himself beyond this with the possibility of selling less than he had planned, with the difficulty of finding potential trading partners or with the possibility that a collapse of credit markets might make it impossible for him to transform future sales into present purchases at any price.

There has been a natural reluctance among most macroeconomists to use a theory based on this conception of ideal coordination for purposes of explaining business-cycle fluctuations, large-scale unemployment and credit crises. These phenomena are obviously characterized by a gross lack of coordination between different agents' economic activities, and by a widespread concern for just those problems which general equilibrium analysis implies can safely be ignored by all agents.

The story of the development of macroeconomic theory beginning with the Keynesian Revolution is largely a story of the struggle between these two opposing forces: the quest for a microfoundation and the recognition that existing micro theory is inadequate for dealing with macro problems. The major innovations in macro theory have consisted of new ways to use the powerful organizing concepts of micro theory, equilibrium and rational choice, to explain phenomena that have traditionally eluded micro theory.

The main analytical innovation of Keynes's *General Theory* was to develop an alternative concept of equilibrium that allowed modified versions of supply and demand analysis to be applied to macroeconomic questions without assuming a state of ideal coordination. The key to this innovation was the recognition that prices were not the only equilibrating variables. In Keynes's equilibrium, the quantity of aggregate output did the equilibrating. Instead of determining employment by the condition that the supply and demand for labour were equal, Keynes imposed the condition that the quantity of output produced equal the quantity demanded, the equilibrium condition of

the familiar Keynesian Cross diagram. Hicks (1937) showed how Keynes's analysis could be formulated as two equations in the two equilibrating variables – output and the rate of interest – a formulation that became the standard paradigm of macro theory for the next 30 years.

With this new concept of equilibrium, Keynesian economics achieved the immediate goal of having a short-run macroeconomic theory with enough equations to determine the variables of interest. Rather than having to treat fluctuations in output and employment as disequilibrium phenomena, using the cumbersome and problematic dynamic techniques available at the time, macroeconomists after Keynes could use the much simpler methods of comparative statics. Furthermore, with this equilibrium concept in hand, they could begin applying choice theory consisted to the analysis of aggregate demand. From Hicks's essay through the 1960s, the main developments in macro theory consisted of the rationalization and modification of the aggregate behavioural relationships postulated by Keynes, through the application of principles of optimization.

Although there was considerable disagreement over whether a Keynesian equilibrium should really be called an equilibrium, and whether it adequately captured Keynes's central ideas, there was a broad consensus among macroeconomists following the Keynesian Revolution concerning the relative domain of applicability of Keynesian macroeconomics and Walrasian micro theory. Modigliani (1944) had shown how Keynesian results could be derived from an otherwise classical model if the money-wage rate were fixed. Since it was widely believed that wages were less than fully flexible in the short run, it seemed natural to see Keynesian theory as applying to short-run fluctuations and general equilibrium theory as applying to long-run questions in which adjustment problems could safely be ignored. This view came to be known as the 'neoclassical synthesis'.

By the 1960s, however, serious doubts were being raised about the logical consistency of this division. Most notably Clower (1965) pointed out that the Keynesian consumption function, a key concept in Keynes's quantity-equilibrating multiplier process, was incompatible with Walrasian general equilibrium analysis. In particular, it was based on the notion that the typical household takes its income, whether current or prospective, as given, whereas in general equilibrium analysis a household is supposed to choose its income, by choosing how many factor-services to sell. Clower raised the question of how a theory with this kind of consumption function could possibly be reconciled with standard microeconomic theory.

The answer proposed by Clower was that Keynesian 'effective' demands would be transmitted by agents in a Walrasian world when the system was not in equilibrium. If general equilibrium prices have not yet been established, then excess demands and supplies will make it impossible for all agents successfully to execute the trading plans that they had formulated on the basis of a single budget constraint. Once they see that this is the case, they will begin to take into account not just their budget constraint but also the quantity limitations implied by non-price rationing. The unemployed worker will base his demands not on the amount of labour he would like to sell at the going wage but on the amount he is selling or expects to sell.

Clower's suggestion was further developed by Barro and Grossman (1971), who also integrated it with Patinkin's (1956, ch. 13) similar analysis of how the demand for labour would be affected by the quantity of output demanded when the system was not in a general equilibrium. Barro and Grossman showed how these ideas could be combined to generalize Keynes's concept of quantity equilibrium. If prices are held fixed at levels that create excess supplies of labour and output, then the equilibrium will generally be a set of quantities that are demanded when agents take into account the sales constraints implied by those quantities.

To many writers, the Barro–Grossman analysis presented a microfoundation for macro theory that confirmed the neoclassical synthesis. Barro and Grossman labelled their contribution a 'general disequilibrium' analysis to emphasize that it generated Keynesian results only if prices were away from their Walrasian equilibrium values. As the

subsequent literature emphasized, this analysis could be combined with the *tâtonnement* mechanism of general equilibrium theory, according to which the price of any good out of equilibrium rises or falls as a function of excess demand or supply of the good. As prices changed, the quantity equilibrium would change with them. The only long-run rest-point to such a system was a Walrasian equilibrium. In the short run, the system would generally be in a Keynesian fixed-price equilibrium.

The major problem with this microfoundation is that it relies on what is generally regarded as the weakest part of micro theory – the *tâtonnement* mechanism. No one has yet successfully integrated that mechanism with the main part of micro theory – the theory of equilibrium. The problem with attempted integrations was posed forcibly by Arrow (1959), who noted that since all agents are assumed to act as price-takers in general equilibrium theory, there is no one who can change prices that are not at equilibrium values. The heuristic device of the 'auctioneer' does more to evade this problem than to solve it.

This problem led several authors in the 1960s and 1970s to turn to the economics of information for a microfoundation. In general equilibrium analysis, the *tâtonnement* can be thought of as the mechanism through which the market collates and disseminates the information required to achieve a coordinated state. When an economy is disturbed by a change in tastes or technology that is at first apparent only to a limited subset of agents, the excess demands and supplies created by this change act as a signal to the rest of society that a changed allocation of resources is called for. The message is passed on to other individuals in the form of a change in relative prices. The neoclassical synthesis pictured macroeconomic problems as arising because this process takes time. The difficulty noted by Arrow was that the informational aspects underlying this process were not present in the decision problems faced by the individuals in the theory. Thus it seemed to many that the way out of the difficulty lay in a more explicit treatment of the role of less than perfect information in individual decision-making.

Considerable progress along these lines was made by various contributors to the famous 'Phelps volume' (Phelps et al. 1970). This volume contained a variety of different approaches to the problem, but the most lasting contribution was the 'island parable' presented in Phelps's introductory essay. According to this parable, the typical transactor trades on a succession of 'informational islands'. Prices on each island always equate demand and supply on that island, but people are unaware of prices and quantities simultaneously prevailing on other islands.

This parable seemed to offer a microfoundation for the neoclassical synthesis without relying on the problematical *tâtonnement* mechanism. In particular, consider an unanticipated purely nominal fall in aggregate demand. According to Phelps's parable, the system would react with a decrease in output and employment and a less than proportional fall in prices in the short run, as in Keynesian theory, but with fully proportional price declines that neutralized any real effects in the long run, as in general equilibrium analysis. The reason for the short-run non-neutrality is that sellers who saw their selling prices fall would tend to read this as a fall in the relative price of their wares, not realizing until later that prices elsewhere in the economy were also falling, and would therefore be induced to sell less. This withdrawal of supply would soften the initial fall in prices. Eventually the realization that this was an aggregate phenomenon, and not just local, would persuade people to supply the same amount as before, and prices would fall all the way to their new equilibrium values.

This apparent microfoundation did not rely on the *tâtonnement* mechanism, but it relied heavily upon the theory of expectation formation. In particular, it postulated that the only impediment to achievement of long-run equilibrium was the slowness with which people formed accurate expectations of the general price level. This postulate left several writers unsatisfied because it implied an incongruity between the formation of trading plans, which agents were assumed to undertake rationally, and the formation of expectations, which they were assumed to do according to a mechanical rule. This dissatisfaction led the

way to the rational expectations revolution in macroeconomics.

The seminal paper in this revolution was Lucas (1972). In this paper Lucas presented an exact model of the island parable, in which agents formed subjective expectations that were the mathematical expectations of the model itself. This expectation scheme was not derived from any explicit optimization scheme; nevertheless, it became known as 'rational' on the belief that people who formed expectations in any other way must be leaving unexploited some opportunities for increasing their well-being. Lucas's model became the analytical paradigm of the school of new classical economics in the 1970s and 1980s, whose research programme was explicitly to base all of macroeconomics upon firm microeconomic principles.

By the early 1980s new classical economics had become the dominant approach in macroeconomic theory. But it was strongly resisted by Keynesian economists, who argued that, although it had firm microfoundations, it was based on a notion of equilibrium that was too close to the frictionless ideal of Walrasian theory. Among other arguments, they objected that the price of avoiding the problems of *tâtonnement* by means of the Phelps island parable was giving up the fundamental Keynesian notion of quantities and other non-price signals as equilibrating variables; hence, giving up hope of explaining many of the obvious coordination problems faced by people in the trough of a business cycle.

In the mid 1980s, however, there has been a resurgence of theoretical support for Keynesian ideas. Specifically, authors like Diamond (1982) and Howitt (1985) have derived models in which all agents are explicitly rational and in which the equilibrium states exhibit Keynesian phenomena. The unifying feature of these models is the assumption that even with perfect price-flexibility, people respond to non-price signals. Specifically, an increase in economic activity on one side of the market (e.g. an increase in aggregate demand) will reduce the costs of trading by making potential trading partners easier to find, and hence will affect the trading decisions on the other side of the market (e.g. will induce an increase in aggregate supply), even if it does not affect market prices. These models are still, however, in their infancy.

It is interesting to speculate on whether or not the quest for a microfoundation will continue to play as important a role in the future development of macroeconomics. The disunity between micro and macro that has motivated so many contributors is shrinking rapidly on the frontiers of research, where micro theory is being transformed by the explicit consideration of informational problems like those so often adduced by macroeconomists, and where macroeconomics without explicit reference to individual transactors, their decision problems and conditions of equilibrium, is becoming increasingly rare.

It is also questionable whether the microeconomic principles of equilibrium and rationality that have been applied so fruitfully in the development of macroeconomics can be of more service. By themselves they are no more than organizing devices; they yield no meaningful empirical propositions in the absence of a great many supporting hypotheses.

M

See Also

- ▶ Equilibrium: an expectational concept
- ▶ New classical macroeconomics
- ▶ Rational expectations

Bibliography

Arrow, K.J. 1959. Towards a theory of price adjustment. In *The allocation of economic resources*, ed. M. Abramovitz et al. Stanford: Stanford University Press.

Arrow, K.J. 1967. Samuelson collected. *Journal of Political Economy* 75(October): 730–737.

Barro, R.J., and H.I. Grossman. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61(1): 82–93.

Clower, R.W. 1965. The Keynesian counter-revolution: A theoretical appraisal. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Diamond, P.A. 1982. Aggregate demand management in search equilibrium. *Journal of Political Economy* 90(5): 881–894.

Hicks, J.R. 1937. Mr Keynes and the 'classics': A suggested Interpretation. *Econometrica* 5(April): 147–159.
Howitt, P. 1985. Transaction costs in the theory of unemployment. *American Economic Review* 75(1): 88–100.
Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2): 103–124.
Lucas, R.E., and T.J. Sargent. 1979. After Keynesian macroeconomics. *Federal Reserve Bank of Minneapolis Quarterly Review* 3(2): 1–16.
Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12(January): 45–88.
Patinkin, D. 1956. *Money, interest, and prices*. New York: Harper & Row.
Phelps, E.S., et al. (eds.). 1970. *Microeconomic foundations of employment and inflation theory.* New York: Norton.

Maddala, G.S. (1933–1999)

Kajal Lahiri

Abstract

G.S. Maddala's many publications covered almost every substantive areas of econometrics – distributed lags, generalized least squares, panel data, simultaneous equations, measurement errors, switching and disequilibrium models, qualitative and limited dependent variable models, selection and self-selection models, exact small sample distributions of estimators, outliers and bootstrap methods, robust estimators and more. G.S. became a veritable textbook himself – a pre-eminent teacher in econometrics and an authority on almost every econometrics topic. G.S. was a brilliant expositor – he could cut through the technical superstructure to reveal only essential details, while retaining the nerve centre of the subject matter he sought to explain.

Keywords

Maddala, G. S; Griliches. Z; Bayesian econometrics; Bootstrap; Cointegration; Crosssection information; Disequilibrium models; Dummy variables; Econometric Society; Econometrics; Elasticity of substitution; Error component estimator; Exchange rate determination; Heteroskedasticity; Instrumental variables; Limited dependent variables; Maximum likelihood; Mundlak, Y; Nerlove, M; Panel data; Pareto distribution; Pascal lag; Production functions; Qualitative variables; Rational distributed lag models; Rational expectations; Recursive models; Sample theory; Simultaneous equations models; Singh–Maddala distribution; Structural breaks; Structural change; Survey data; Technical change; Time series analysis; Two-step procedures; Unit roots; Wald test; Weibull distribution

JEL Classifications

B31

G.S. Maddala (universally known as 'G.S.') was born on 21 May 1933 in the south Indian state of Andhra Pradesh, where he had his high-school education. G.S. held the University Eminent Chair at the Ohio State University when he died on 4 June 1999 due to congestive heart failure.

G.S.'s father was a schoolteacher of modest means, and his mother, though having only an elementary education, was well versed in Sanskrit and the works of the great Indian philosopher Sankara. After graduating from high school in 1947, G.S. had to drop out of college for a few years due to health and other reasons. In 1955 he graduated first in his class from Andhra University with a BA in mathematics, and went on to graduate in First Class from Bombay University with an MA in statistics in 1957. With a Fulbright Fellowship, G.S. travelled to the University of Chicago in 1960 and completed his Ph.D. in 1963 under the supervision of the late Zvi Griliches. In that year, he was offered the job of Assistant Professor of Economics at Stanford University. Before joining Ohio State in 1993, G.S. taught at the University of Rochester (1967–1975) and at the University of Florida (1975–1993). He also held visiting appointments at Cornell, Yale, CORE, Monash, Columbia, Caltech (as the Fairchild Distinguished Scholar), Emory and Oakridge Labs. The fascinating narration of his journey from an early college dropout

in a remote Indian village in 1947 to a faculty position at Stanford in 1963 can be found in the Introduction ('How I Became an Econometrician') to the two-volume selected works of Maddala (1994). More detailed biographical information, his life story and philosophy can be found in Lahiri and Phillips (1999), Lahiri (1999), Griliches (1999), Rosen (2000) and Hsiao (2003).

Beginning with his first published paper (with Zvi Griliches, Robert Lucas, and Neil Wallace) in 1962, through the next four decades, G.S. published 12 books and more than 110 articles covering almost every emerging area of econometrics – distributed lags, generalized least squares, panel data, simultaneous equations, measurement errors, tests of significance, switching and disequilibrium models, qualitative and limited dependent variable models, selection and self-selection models, exact small sample distributions of estimators, outliers and bootstrap methods, robust estimators, and more. The list is practically endless. Throughout his career G.S. used sample theory and Bayesian techniques freely in his research, a rarity in the econometrics profession, and was one of the early proponents of Bayesian approach in econometrics. Through his many books and the breadth of his own research, G.S. became a veritable textbook himself – a pre-eminent teacher in econometrics and an authority on almost every econometrics topic. Not surprisingly, according to the *Social Science Citation Index*, G.S. was one of the top five most-cited econometricians during each of the years 1988–1993, and he was cited more times in 1994 and 1996 than each of the six econometricians who won the Clark Medal during 1970–2000.

During the 1960s, G.S. contributed heavily towards the formulation and estimation of production functions and technical change. His doctoral dissertation was on productivity and technical change in the US bituminous coal industry. His two papers with Jay Kadane in 1966 and 1967 considered, respectively, the importance of alternative exogeneity assumptions in the estimation of the constant elasticity of substitution production functions parameters inclusive of the share equations; and the bias in the estimation of the returns to scale parameter when the production function is incorrectly specified as a Cobb Douglas. The rigour and depth in these papers were undoubtedly ahead of their time.

The early 1970s saw a flurry of activity on efficient estimation methods of alternative distributed lag models. One of G.S.'s widely cited papers (1971a) showed why certain commonly used two-step procedures are asymptotically less efficient than the maximum likelihood estimator in the presence of lagged dependent variables as regressors. This sort of problem is encountered also in dynamic panel data models with individual heterogeneity. The key result in this paper is that in these models the information matrix of the slope parameters and the parameters embedded in the covariance matrix of residuals are not diagonal. Using this as a starting point, Pagan (1986) developed a more thorough and modern characterization of numerous two-step procedures with estimated covariance matrix in the context of various econometric models.

With Dave Grether in 1973, G.S. studied the effects of errors in variables in distributed lag models with serial correlation. They showed analytically that the estimated speed of adjustment can be severely biased, and can give the spurious appearance of a long lag in adjustment. In two influential papers with A.S. Rao in 1971 and 1973, G.S. developed maximum likelihood procedures for Solow's Pascal lag and Jorgenson's rational distributed lag models, and compared the power of tests for serial correlation in regression models with lagged dependent variables. One important conclusion that emerged from the latter study was that the nature of the autocorrelation and trend in the exogenous variable is crucial in determining the small sample behaviour of the test statistics and the estimators – hinting at much of the work on integrated variables that would come in the 1980s.

During the early 1970s G.S. also produced a number of important papers on the use and estimation of panel data models, and rightfully became one of the three 'fathers' (together with Yair Mundlak and Marc Nerlove) of modern panel data analysis in econometrics. In his influential *Econometrica* (1971b) paper, G.S. demonstrated – with his characteristic clarity –

that the error component estimator is a weighted combination of within and between estimators, and thus the use of dummies entails substantial loss of information by ignoring the 'between' variation in the data. In another *Econometrica* (1971c) paper, G.S. discussed the problem of pooling cross-section and time series data, and emphasized tests for consistency between time series and cross-section information. The paper contains a very deep analysis of an alternative Bayesian approach with diffuse priors and concludes that the two approaches should be complementary. (Publishing three full-length articles in *Econometrica* in a year has to be some kind of a record for an economist!) The profession quickly saw the enduring value of these publications and elected G.S. a fellow of the Econometric Society in 1975.

During the 1970s, like many other econometrics stalwarts of the period, G.S. was also involved in the development of econometric methodology in simultaneous equations models. He worked on appropriate estimation strategies in large and medium-size econometric models (1971d), and studied the power characteristics of alternative tests of significance associated with simultaneous equation estimation (1974a). His *Econometrica* (1974b) paper showed that 'diffuse' and 'non-informative' priors might lead to sharp posterior distributions even in under-identified models. Only recently have Chao and Phillips (2002) fully solved the so-called 'Maddala paradox' using Jeffreys prior. They interpret the pathological result in terms of a naive use of the diffuse prior that fails to downweight sufficiently that part of the parameter space where the rank condition either fails or nearly fails. In another potent contribution to an important recent work on weak instruments, Maddala and Jeong (1992) correctly showed that the bimodal distribution of the instrumental variable estimator obtained in the literature is merely due to the illustrative model used, where the correlation between the structural and the first-stage errors is perfect. Phillips (2006) gives a complete characterization of the bimodality problem when instruments are weak.

From the mid-1970s, G.S. was primarily focused on developing estimation and test procedures for qualitative and limited dependent variable models, and produced nearly 40 articles. This line of research also dealt with models with selection, self-selection, disequilibrium and controlled prices. His work at Rochester with Forrest Nelson (1974) on disequilibrium models and with Lung-Fei Lee (1976) on recursive models with qualitative endogenous variables and generalized selection models represents a long and very fruitful period of research on this topic. His 1983 Econometric Society monograph, *Limited Dependent and Qualitative Variables in Econometrics*, was an immediate best-seller and was declared a citation classic in *Current Contents* (vol. 30, 16 July 1993). It has fuelled much of the innovative applied and theoretical research using these tools since the mid-1980s, and has served as a bible to empirical researchers in applied microeconomics. The strength of the book lies in its comprehensiveness, expositional simplicity, and depth. As of June 2006, the Google Scholar reports a record 3,721 citations of this advanced monograph. G.S. also wrote a number of theoretical and empirical papers analysing limited dependent and qualitative variable models with panel data, and wrote widely cited expository articles for use in other disciplines such as accounting, finance, transportation, and health.

It is notable that G.S. can jointly claim a statistical distribution – the Singh–Maddala (1976) distribution – a much better name than the Burr type 12 to which it is related. Maddala and Singh's proposed statistical distribution has triggered much research in describing the actual size distribution of incomes, and is a generalization of the Pareto distribution and the Weibull distribution used in analysis of equipment failures. As aptly noted by Sherwin Rosen (2000) while delivering the first Maddala lecture at Ohio State University on 26 April 2000, 'Coase may have his Theorem, Stigler his Laws, Black and Scholes their Formula, and Lucas his critique, but what economist aside from Pareto (who was just as much a sociologist and political scientist and only one third economist) has half ownership of a distribution? And what an elegant economic derivation it has.'

G.S. had a deep interest in rational expectations models, in the validity of the hypothesis that

can be gleaned from recorded survey data, and in how econometric disequilibrium models play out in this framework. Maddala, Fishe and Lahiri (1983) developed methods to estimate aggregate expectations when available survey data are partly qualitative and partly quantitative. He had done pioneering work (Maddala, 1983a) on the estimation for models with bounded price variation, and with Scott Shonkwiler (1985) applied the methodology to the corn market. With Steve Donald (1992), G.S. studied the disequilibrium model with upper and lower bounds on prices under rational expectations. The latter paper foreshadowed much work on exchange rate determination in a target zone in the 1990s. Undoubtedly, the full potential of this line of research initiated by G.S. is yet to be realized.

With failing health, G.S. spent much of the 1990s working primarily on bootstrap techniques and time series models with cointegration and structural breaks. During this period, he also wrote important papers on tests of unit roots in panel data models, robust inference, errors in variables problems in finance, Bayesian shrinkage estimation, outliers and influential observations, neural nets, and many others. Thus, ill health neither slowed down his research nor dampened his passion for mentoring and supervising Ph.D. students. In total G.S. supervised close to 60 doctoral students, co-authoring more than 65 published articles with them.

While testing the rationality of survey data on interest rate expectations in the context of a multiple-indicator single index model with heteroskedasticity, Maddala and Jeong in the mid-1990s used the weighted double bootstrap method to implement the Wald test in finite samples. His work with Hongyi Li in 1996 explored the use of different bootstrap techniques in cointegration regressions, financial and non-linear models. With Wu (1999) on panel data unit root test, G.S. suggested the use of a novel Fisher test that combines N individual tests with bootstrap-based critical values. Since much remains to be done to extend the Fisher approach to combining individual tests that are correlated, further generalizations of the Maddala–Wu test are certainly to come.

Much of his work on modern time series analysis has been summarized in his seminal book with In-Moo Kim (1998). This book also presents a comprehensive and lucid review of unit root and cointegration tests, and estimation with integrated variables. It discusses problems of unit root tests and cointegration under structural change, outliers, robust methods, the Markov switching model, and Harvey's structural time series model. The book contains a welcome chapter on the Bayesian approach to many of these problems and bootstrap methods for small-sample inference.

G.S. contributed to a number of purely policy-oriented and applied areas. Some of these topics include consumption, production and cost functions, money demand, regulation, pseudo-data, returns to college education, housing markets, energy demand, stock prices, international macro, and cross-country growth analysis. In all these papers, G.S. made serious attempts to grapple with substantive and important issues of the day. However, one common characteristic that flows through all these papers is that they unfailingly reflect the discriminating judgement of a consummate econometrician.

G.S. had the gift of a brilliant expositor – the ability to cut through the technical superstructure to reveal only essential details, while retaining the nerve centre of the subject matter he sought to explain. He loved to write econometrics in plain English. There was magic in how he could cut to the core, strip away all the irrelevant details and illuminate the essence of the issue in a quiet and unassuming way. This exceptional expository capability made him revered by applied and theoretical econometricians alike. This skill was apparent in all his writing and was a central element in his textbook expositions. His 1977 econometrics text redefined the boundaries of econometrics that could be integrated into graduate teaching, and became a new standard for subsequent econometrics textbooks. His advanced undergraduate textbook *An Introduction to Econometrics* has gone into its third edition (2000), and all his textbooks have been translated into a number of foreign languages.

G.S.'s style was to take a critical but constructive look at evolving econometric techniques – in

particular those that have little practical significance. In this, G.S. had something that was close to perfect pitch in econometrics. He was one of the few econometricians who constantly asked whether the questions being answered were worth asking – always maintaining a clear perspective on a wide range of issues in econometrics and their relationship to economic problems. In doing so, he never hesitated to go against the tide of the profession. While much of his work was undoubtedly constructive, much was also critical of many current fads in econometrics. That is also a very important contribution.

See Also

- Bayesian Econometrics
- Bootstrap
- Categorical Data
- Distributed Lags
- Econometrics
- Elasticity of Substitution
- Expectations
- Fixed Effects and Random Effects
- Inequality (Measurement)
- Rational Expectations
- Roy Model
- Selection Bias and Self-selection
- Shrinkage-biased Estimation in Econometrics
- Time Series Analysis

Selected Works

1962. (With Griliches, Z., R.E. Lucas, and N. Wallace.) Notes on estimated aggregate consumption functions. *Econometrica* 30: 491–500.

1965. Productivity change in the bituminous coal industry. *Journal of Political Economy* 73: 352–365.

1966. (With Kadane, J.R.) Notes on the estimation of elasticity of substitution. *Review of Economics and Statistics* 48: 340–344.

1967. (With Kadane, J.R.) Estimation of returns to scale and elasticity of substitution. *Econometrica* 35: 419–423.

1971a. Generalized least squares with an estimated covariance matrix. *Econometrica* 39: 23–33.

1971b. On the use of variance component models in pooling cross-section and time series data. *Econometrica* 39: 341–358.

1971c. The likelihood approach to pooling cross-section and time series data. *Econometrica* 39: 939–953.

1971d. Simultaneous equations methods for large and medium-sized econometric models. *Review of Economic Studies* 38: 435–445.

1971. (With Rao, A.S.) Maximum likelihood estimation of Solow's and Jorgenson's distributed lag models. *Review of Economics and Statistics* 53: 80–88.

1972. (With Grether, D.M.) On the asymptotic properties of two-step procedures used in the estimation of distributed lag models. *International Economic Review* 13: 737–744.

1973a. (With Grether, D.M.) Errors in variables and serially correlated residuals in distributed lag models. *Econometrica* 41: 255–262.

1973b. (With Rao, A.S.) Tests for serial correlation in regression models with lagged dependent variables and serially correlated errors. *Econometrica* 41: 255–262.

1974a. Some small sample evidence on tests of significance in simultaneous models. *Econometrica* 42: 841–851.

1974b. Weak priors and sharp posteriors in simultaneous equation models. *Econometrica* 44: 345–351.

1974c. (With Nelson, F.D.) Maximum likelihood methods for models of markets in disequilibrium. *Econometrica* 42: 1013–1030.

1976a. (With Lee, L.-F.) Recursive models with qualitative endogenous variables. *Annals of Social and Economic Measurement* 5: 525–545.

1976b. (With Singh, S.K.) A function for size distribution of incomes. *Econometrica* 44: 963–970.

1977. *Econometrics*. New York: McGraw Hill.

1983a. Methods for models of markets with bounded price variation. *International Economic Review* 24: 361–378.

1983b. *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.
1983c. (With Fishe, R.P., and K. Lahiri.) A time series analysis of popular expectations data. In *Economic applications of time-series analysis*, ed. A. Zellner. Washington, DC: U.S. Census Bureau.
1985. (With Shonkwiler, J.S.) Modeling expectations of bounded prices: An application to the market for corn. *Review of Economics and Statistics* 67: 697–702.
1992a. (With Jeong, J.) On the exact small sample distribution of the instrumental variable estimator. *Econometrica* 60: 181–183.
1992b. (With Donald, S.) A note on the estimation of limited dependent variable models under rational expectations. *Economics Letters* 38: 17–23.
1994. *Econometric methods and applications: Selected papers of G.S. Maddala,* volume 2. Aldershot: Edward Elgar.
1996. (With Li, H.) Bootstrapping time series models (with discussion). *Econometric Reviews* 15: 115–195.
1998. (With Kim, I.-M.) *Unit roots*, *cointegration and structural change*. Cambridge: Cambridge University Press.
1999. (With Wu, S.) A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics* 61: 631–652.
2000. *An introduction to econometrics*, 3rd edn. Chichester: Wiley.

Bibliography

Chao, J.C., and P.C. Phillips. 2002. Jeffreys prior analysis with simultaneous equations models in the case with n +1 endogenous variables. *Journal of Econometrics* 111: 251–282.
Griliches, Z. 1999. Forward. In *Analysis of panels and limited dependent variable models* (essays in honor of G.S. Maddala), ed. C. Hsiao et al. Cambridge: Cambridge University Press.
Hsiao, C. 2003. In Memoriam: G.S. Maddala. *Econometric Reviews* 22: vii–viii.
Hsiao, C., K. Lahiri, L.-F. Lee, and M.H. Pesaran. 1999. *Analysis of panels and limited dependent variable models* (essays in honor of G.S. Maddala). Cambridge: Cambridge University Press.
Lahiri, K. 1999. ET interview: Professor G.S. Maddala. *Econometric Theory* 15: 753–776.
Lahiri, K., and P.C. Phillips. 1999. Obituary: G.S. Maddala, 1933–1999. *Econometric Theory* 15: 639–641.
Pagan, A.R. 1986. Two stage and related estimators and their applications. *Review of Economic Studies* 53: 517–538.
Phillips, P.C. 2006. A remark on bimodality and weak instrumentation in structural equation estimation. *Econometric Theory* 22: 947–960.
Rosen, S. 2000. G.S. Maddala Memoir. First G.S. Maddala Lecture. Department of Economics, Ohio State University, Columbus, 26 April 2000.

Mahalanobis, Prasanta Chandra (1893–1972)

Sukhamoy Chakravarty

M

Keywords

India, economics in; Mahalanobis, P. C.; Optimal growth; Planning; Prebisch, R.; Statistical inference; Two-sector models; Fel'dman, G. A.; Dual development thesis

JEL Classifications

B31

Mahalanobis was born in Calcutta of a well-to-do Bengali middle-class family with a reformed outlook on Hindu religion. He was educated first at Presidency College, Calcutta, and then at Cambridge, where he graduated with a First in Natural Sciences from King's College in 1915. He became a Fellow of the Royal Society in 1946 and received many other scientific honours. While Mahalanobis served as Professor of Physics at Presidency College for nearly three decades, his scientific work consisted chiefly of developing statistical theory and techniques that had

I am grateful to Anthony Davies, Cheng Hsiao, Kay, Tara and Vivek Maddala, Thad Mirer, Peter Phillips and others who have contributed to this biography

application to a wide range of subjects, beginning with meteorology and anthropology and ending in economics.

Mahalanobis established a firm international reputation on the basis of his work on the design of large-scale sample surveys (for example, 1944) and thus laid the basis for systematic collection of a large variety of data relating to socio-economic conditions. Mahalanobis's sense of realism was combined with a deep understanding of the problems of statistical inference. This led him to place stress on 'non-sampling errors' in addition to the standard preoccupation with sampling errors. He devised his system of 'interpenetrating network of sub-samples' to derive among other things, an idea of 'non-sampling errors' which are inherently associated with large-scale collection of data.

Mahalanobis's work on experimental designs developed with a view to estimating crop yields (1946) was highly influential in laying down the basis for collection of agricultural statistics in India. In multivariate statistics, Mahalanobis's measure of distance between two populations (1936), usually known as Mahalanobis's D^2 statistic, is a major contribution that is much used in anthropometry and elsewhere.

Mahalanobis maintained a keen interest in problems of national planning even before India had gained Independence. He recognized very early that such planning had to have a firm statistical base, and from the beginning of the 1950s, when the Indian Five Year Plan was launched, began to devote a very large part of his time and attention to questions of estimating national income and the factors determining its rate of growth. His approach to planning issues, with its strong emphasis on quantification, was significantly different from the qualitative approach favoured by the Indian economists of his generation. However, Mahalanobis was no exclusive believer in narrowly conceived quantitative techniques. He developed an important blend between qualitative and quantitative considerations, which is reflected in his 'Approach of Operational Research to Planning in India' (1955).

The second Five Year Plan, whose analytical structure was largely the handiwork of Mahalanobis, stands out as a very distinguished document in the development of planning theory. Mahalanobis is generally regarded as one of the prominent advocates of the inward-looking strategy of industrialization, along with Raul Prebisch. But the analytical foundation of the Mahalanobis approach was derived from somewhat different premises. While Prebisch began his theoretical study from what he thought was a historical fact, that is, the secular decline in terms of the trade of primary producing countries, Mahalanobis developed a two-sector model of growth to deduce a strategy of industrial development which he thought was best suited to India. The classification of the economy into sectors resembled in some respects Marx's famous Departmental Schema, although they were not identical.

Mahalanobis's sector-schema (1953) distinguished between 'capital goods' and 'consumer goods', but the assumption of vertical integration made in the interest of simplicity made statistical implementation difficult. The essential point of the model is that the capacity of the capital goods sector determines the potential rate of expansion of the consumer goods sector, and not the other way round. Further, at any given instant, capacities are not directly transferable from one sector to the other. Labour is not considered to be a constraint on expanding production. The model was developed initially for a closed economy but has been subsequently extended to open economies, with an exogenously given profile of export earnings. Mahalanobis used the model to illustrate the nature of the trade-off between present and future consumption, given the objective characteristics of the two sectors.

For the dynamic closure of the model he used the ratio of the output of the capital goods sector that is ploughed back into itself ('λ_k' in his notation), to deduce a 'gradualist growth' path of consumption. For any given value λ_k maintained over time, the rate of growth of aggregate output tends, over a sufficiently long period, to a magnitude $\lambda_k\beta_k$, where β_k is the output–capital ratio of the capital goods sector. The Mahalanobis model was subsequently freed from the assumption of an exogenously stipulated λ_k Exercises carried out by Stoleru (1965), Chakravarty (1969),

Dasgupta (1969) and others introduced explicit intertemporal social utility functions along with a production technology of the Mahalanobis type. They deduced the characteristics of optimal growth paths with the help of variational calculus. $\lambda_k(t)$ was deduced as a solution of the optimizing exercise. It was shown that while the assumption of 'non-shiftability' critical to Mahalanobis's model could in several cases give rise to a preference for capital goods sector in early stages of growth (a strategy preferred by Mahalanobis himself), one could not obtain a universal rule of priority for capital goods irrespective of initial conditions, or the nature of social utility functions over time.

In all these exercises, the coefficients pertaining to the 'capital goods sector', sometimes identified as the 'machine tool sector', turned out to be an important determinant of the growth process. Earlier literature on business cycle theory originating with Marx, Tugan Baranovsky and Adolph Lowe had placed emphasis on the 'machine tools sector', without linking it up with an explicit growth model. In the growth-theoretic area Fel'dman alone appears to be the true predecessor of Mahalanobis, as is evident from Domar's discussion (1957).

Mahalanobis extended the two-sector model to a four-sector model, to focus on issues of reduction in unemployment along with increases in income. Mahalanobis came to the 'dual development thesis', which consisted in assigning high weights to the capital goods sector in the interests of long-term growth, and emphasis on the highly labour-intensive consumer goods sector in the short run. In the literature on planning, this has on occasion been referred to as the strategy of 'walking on two legs', with authorship occasionally ascribed to Mao Tse Tung.

Towards the end of his life, Mahalanobis returned to issues of statistical methodology and concentrated on developing what he called 'fractile graphical analysis' (1960), which is based on a geometrical concept of error and can also provide a generalized measure of separation between two 'different universes' of study.

Mahalanobis's work remains important for economists who are working on quantitative approaches to problems of plan formulation, especially in the context of large-sized economies. His work on sample surveys has generated a very valuable literature to which economic statisticians from India and elsewhere have made notable contributions.

See Also

- ▶ Development Economics
- ▶ Fel'dman, Grigorii Alexandrovich (1884–1958)

Selected Works

1936. On generalized distance in statistics. National Institute of Science, India. *Proceedings* 2: 49–55.
1944. On large scale sample surveys. In *Philosophical transactions,* Series B, vol. 231, 392–451. London: The Royal Society.
1946. Sample surveys of crop yields. *Sankhya* 7: 269–280.
1953. Some observations on the process of growth of national income. *Sankhya* 12: 307–312.
1955. The approach of operational research to planning in India. *Sankhya* 16(1–2): 3–130.
1958. Science and national planning. *Sankhya* 20: 69–106.
1960. A method of fractile graphical analysis. *Econometrica* 28: 325–351.
1961. *Talks on planning*. Bombay/London: Asia Publishing House.

Bibliography

Bhagwati, J.N., and S. Chakravarty. 1969. Contributions to Indian economic analysis – A survey. *American Economic Review* 59 (4, Supplement): 1–73.
Chakravarty, S. 1969. *Capital and development planning*. Cambridge, MA: MIT Press.
Dasgupta, P.S. 1969. Optimum growth when capital is non-transferable. *Review of Economic Studies* 36 (1): 77–88.
Domar, E.D. 1957. *Essays in the theory of economic growth*. New York/Oxford: Oxford University Press.
Rao, C.R. 1973. Prasanta Chandra Mahalanobis. In *Biographical memoirs of the fellows of Royal Society*, vol. 19. London: The Royal Society.
Stoleru, L.G. 1965. An optimal policy for economic growth. *Econometrica* 33: 321–348.

Maine, Sir Henry James Sumner (1822–1888)

O. Kurer

Maine studied classics at Cambridge, was Professor of Civil Law at Cambridge (1847–54) and was appointed Reader at the Inns of Court in London (1852). The years from 1862 to 1869 he spent as Legal Member of the Council of India. After his return he became Professor of Law at Oxford (1869) and Cambridge (1887) and was elected Master of Trinity Hall, Cambridge in 1877. In addition, Maine was an active journalist.

Maine's most important work is *Ancient Law* (1861), where he introduced the historical method of jurisprudence to an English audience. He is best known today for his proposition of the 'movement of progressive societies from Status to Contract'. In early society, man's rights and duties were rigidly determined by his position at birth by his parents' position. In modern society, voluntary contractual connections predominate. Of particular interest to him was the mechanism by which legal change occurs: he postulated a progression from legal fictions (pretending there is no change) to equity (exceptions in particular cases) to direct changes of the law by virtue of the authority of power.

Maine was a resolute opponent of democracy. Progress in his view was achieved by enlightened minorities, the extension of democracy to the mass of men threatened stability and progress. He strongly resented interference with the freedom of contract, a retrogressive step according to his view of evolution.

Maine greatly influenced the English historical school, in particular through Cliffe Leslie. With them he believed that scientific historians were in a position to discern clues as to the processes governing historical change.

See Also

► Historical Economics, British

Selected Works

1861. *Ancient law: Its connection with the early history of society, and its relations to modern ideas*. London: Murray.

1871. *Village communities in the East and West, to which are added other lectures, addresses, and essays*. London: Murray.

1885. *Popular government*. London: Murray.

Bibliography

Bock, K.E. 1974. Comparison of histories: The contribution of Henry Maine. *Comparative Studies in Society and History* 16: 232–262.

Burrow, J.W. 1968. *Evolution and society. A study in Victorian social theory*. Cambridge: Cambridge University Press.

Feaver, G.A. 1969. *From status to contract. A biography of Sir Henry Maine, 1822–1888*. London: Longmans.

Grant Duff, M.E. 1892. *Sir Henry Maine: A brief memoir of his life*. New York: Holt.

Smith, B.C. 1963. Maine's concept of progress. *Journal of the History of Ideas* 24: 407–412.

Maintaining Capital Intact

K. H. Hennings

The purpose of economic activity is normally to produce a surplus which can be distributed (or appropriated) as income. But only that part of total output can be reckoned as surplus and distributed as income which is not needed to provide for the continuation of economic activity on the same level. Thus in an economy producing only corn (from abundant land and labour) an amount of total output equal to what was required to produce it has to be set aside as seed, and only the remainder can be distributed as income and is at the most available for consumption. If more is distributed, the economy cannot continue on the same level as before, that is, becomes less viable, and will ultimately become unviable so that economic activity comes to an end. To set aside as

much output as is required to continue economic activity on the same level as before (or, to provide for simple reproduction) is thus a principle which may be violated in the short run, but which endangers the viability of economic activity if not adhered to in the long run. If economic activity involves stocks which are considered as 'capital', this principle requires that such capital be maintained intact. Capital is maintained intact if an amount of total output is set aside for its maintenance which ensures that the remaining output is on a level that can be kept up forever. The principle that the viability of economic activity be maintained which requires that capital be maintained intact, thus implies the Hicksian income concept (Hicks 1936, ch. 14).

This formulation of the principle assumes that it is desirable to maintain the current level of economic activity. While it is reasonable to assume that it is not desirable to 'eat up capital' over more than a short period (despite exceptions such as wars), it is often considered desirable to consume less than the maximal possible amount, i.e. to accumulate capital and thus to increase the level of economic activity that can be permanently maintained. The principle that the viability of economic activity be maintained therefore sets a minimum standard. By comparison with propositions about the optimum accumulation of capital such as the golden rule, it is a proposition about the minimum maintenance of capital.

As here formulated the principle is empty as long as what it means to maintain capital intact is not defined. How this should be done has been the subject of long-standing debates among accountants and business economists in the guise of determining the correct principles for drawing up business accounts and calculating business income. Much the same issues have been discussed by social accountants from a macroeconomic perspective. In economic theory they have been the subject of an important debate between A.C. Pigou and F.A. Hayek in the 1930s and 1940s.

It will be useful to begin with the accountants' perspective. It has often been maintained, reasonably enough in view of the legal obligations of a firm to its creditors, that business accounts should be drawn up such that the amount of money capital invested in the firm be preserved, and that business income should be reckoned as distributable only after allowance has been made for its maintenance. This raises a host of valuation difficulties (see Parker and Harcourt 1969), but the principle is reasonably clear: what should be preserved is the net present value of the firm, corrected for price level changes, i.e. measured in real terms. That is what matters from the point of view of someone who has invested in the firm. It is obvious that this implies that assets are valued not at historical cost, nor at replacement cost, but rather according to their most probable expected future contribution to business income. That means taking account of such factors as business risks as well as planned obsolescence when the expected physical life of an asset is longer than its expected use. Nor does preserving the net present value of a firm imply that any particular asset be preserved, or maintained intact. It may be more profitable to run down an asset deliberately, i.e. shorten its physical life (and even productivity) by neglecting to keep it in good working order, or even get rid of it altogether. For what is to be preserved is not capital in the sense of a stock of physical objects (capital goods), but net wealth. In an accountant's perspective, maintaining capital intact means preserving net wealth, and not a stock of capital. This can be done in a variety of ways, including a change in the stock of capital. Hence there is no necessary connection between preserving net wealth and maintaining a stock of capital, from the point of view of a firm. The same holds true for any other microeconomic unit, in particular households. In a microeconomic perspective, therefore, income is whatever is earned minus what is necessary to preserve, given current expectations, the net present value of the firm or household in real terms.

In a macroeconomic perspective, economic activity can without loss of generality be considered as production which involves capital goods such that at any moment there exists a stock of capital. Consider an arbitrary time period. Both at the beginning and the end of the period there will exist capital stocks which consist of different types of capital goods of different ages. Some of

the capital tools existing at the beginning of the period will have been used up, or have been discarded, during the period. Others will have been used during the period, and will normally have been subject to wear and tear; whether they have been maintained in good working order or not, they are all older at the end of the period. Finally, there will be new capital goods, produced during the period. So in general the capital stock existing at the end of the period will consist of different capital goods than those which formed the capital stock at the beginning of the period. However, if the economy is in a stationary state, the size and composition of the capital stock in the same at the beginning and the end of the period. As the age structure of all types of capital goods will be constant, replacing the oldest items of each type of capital good will ensure that all losses are replaced and all wear and tear is made good. Where that is not possible because the period is too short, and the durability of capital goods so long that none is due to be replaced, a fund can be set up to accumulate what is required for replacement. In such a stationary situation no investment decisions need to be taken. Because the structure of production is well integrated, all types of capital goods needed for replacement are produced, and can be used for the purpose for which they were produced. Similarly, all consumption goods will be used for consumption, and will together constitute the income of the economy.

Much the same will be true in an economy growing in a steady state as long as there is no technical change. There will be investment, but investment decisions will be confined to decisions to produce more of everything. The capital stock at the end of the period will be larger than that at the beginning, but its composition will be the same: all outgoing capital goods will simply be replaced by a larger amount of the same type. Income will consist of all consumption goods and the excess of new capital goods over what would have been required in a stationary situation.

In both these cases the meaning of 'maintaining capital intact' is clear. In a stationary economy it means keeping the size and the composition of the capital stock constant. In a steady state without technical change it means letting the capital stock grow while keeping its composition constant. In all other situations, i.e. if there is technical change or the economy is in a non-steady state, both the size and the composition of the capital stock will change, and maintaining capital intact can no longer be defined with respect to identical types of capital goods. Hence it was sought to specify 'maintaining capital intact' in a different manner.

Pigou (1932, 1935) proposed that 'maintaining capital intact' should be defined in such a way that a capital stock should be considered as constant if the capital goods which were used up or discarded during the period are replaced by other the production of which requires an amount of real resources equal in that period to what would have been required to reproduce the same types of capital goods as the used up or discarded ones. Because he supposed that the replacements are 'so chosen that the maximum possible addition is made to the present value of the stock of capital' (1935, p. 239), Pigou considered the capital stock as 'an entity capable of maintaining its quantity while altering its form and by its nature always drawn to those forms on which, so to speak, the sun of profits is at the time shining' (1935, p. 239). Maintaining capital intact thus meant to Pigou the replacement of outgoing items by items with the same real cost of production. It did not include provisions for making good wear and tear. Moreover, Pigou distinguished 'normal' maintenance form replacements made necessary by capital losses due to 'the act of God or the King's enemies', or acts of Parliament, which are not covered by consideration of the risks attached to the normal conduct of business (1935, p. 240). Real income, in his conception, was then total output minus what was required for 'normal' maintenance.

Hayek (1935, 1941a, b) criticizes Pigou's proposals on several grounds. First of all Hayek objected to the backward-looking nature of Pigou's capital concept and pointed to the absurdities it could entail. Instead, he insisted on a forward-looking concept in which capital goods are valued (in real terms) on the basis of their prospective quasi-rents and not their costs of production. This point Pigou accepted, and later

defined a capital stock to be constant if the items which replace outgoing capital goods are 'expected to yield the same income' as those they replace (1941, p. 274). Hayek also objected to Pigou's failure to include the making good of wear and tear in what was required to maintain capital intact, and argued at the same time that in a changing economy (which is subject to technical change, or in a non-steady state, or both) obsolescence is far more serious than physical wear and tear. Where obsolescence is planned, or expected, not to include allowances for it would underestimate what is required to maintain capital intact, and overestimate income. Where obsolescence is unexpected, attempts will be made to adapt as far as possible capital goods which were produced for one purpose to other uses. Not to make allowance for such attempts, or to do so only when such capital goods are finally discarded and replaced (as Pigou's proposal implies), would equally distort the calculation of both maintenance and income. Hayek also pointed out that all expected changes in the prospective quasi-rents of capital goods, say as a result of technical progress, or of unexpected price changes, represent capital gains or losses in real terms. Such windfalls should in his view affect maintenance rather than income. Hayek emphasized strongly that maintaining capital intact was not an end in itself, as Pigou seemed to imply, but a means to achieve a constant flow of income in order to avoid what D.H. Robertson (1933) had called unintended stinting or splashing, i.e. consuming either too little or too much to maintain the present level of consumption. Hence he argued that windfall gains and losses should affect income only in so far as they can be converted into permanent increases or decreases in the flow of income.

Hayek's main point was thus that in a changing economy changes in prices and expectations will lead to both real and price Wicksell effects, i.e. changes in the methods of production and in the valuation of capital goods. Both will cause entrepreneurs to adapt their investment behaviour because, aiming at a constant flow of income, they will not allow windfall capital gains and losses to affect their income beyond what can be converted into permanent increases and decreases in income. Hence precisely because entrepreneurs aim at a constant flow of income they will again and again adapt the capital stock required to produce that flow. There is thus no reason why the capital stock should in some sense be constant. Moreover, changes in capital stocks between the beginning and the end of a period reflect not so much physical wear and tear as obsolescence and adaptations due to changed circumstances and expectations. Hence maintenance cannot be distinguished from net investment, and attempts to define income by taking account of what is required to maintain capital are bound to fail. Hayek's argument thus amounts to a demonstration that in a changing economy, maintaining capital intact in the sense of keeping a stock of capital constant in some sense or another was incompatible with the Hicksian income concept, i.e. with maintaining income constant. Pigou (1941, 1946) responded to Hayek's capital theoretic critique by restating his position. Writing from the point of view of welfare theory, and dealing with what would now be called ex post aggregates, Pigou insisted that some method, however rough and ready, of separating maintenance from new investment should be found. That is what Hicks (1942) attempted in his contribution to the debate. While accepting Hayek's critique, he doubted the appropriateness of maintaining a constant flow of income in a changing economy, and at the same time proposed a method, based on Lindahl (1933), designed to separate ex post the consequences of expected from unexpected changes in the value of capital goods, so that the former could be charged to maintenance, and the latter as windfalls to income (see also Hicks 1958, 1969, 1973, ch. 13; and especially 1979). More recently, Scott (1984) has made a similar attempt, placing the dividing line between depreciations and appreciations due to relative price changes which should fall on income, and other changes which should fall on maintenance.

Despite some constructive efforts Hayek's critique is essentially negative. As Hicks (1974) has pointed out, Pigou's position can be associated with the attempt to provide a concept of a capital stock which corresponds to the flow of income it helped to produce. From the point of view of the

construction of a production function, such an attempt is valid; but it does come up against the twin problems of technical progress and non-stationary conditions. So the problem which Pigou originally attempted to answer is still unsolved. Progress may be expected from extending recent advances in capital theory, which so far have been confined to comparative static analysis, to the analysis of a changing economy. A beginning has been made with the development (Hicks 1973) of the concept of a 'transition' from one steady state to another, but more work needs to be done. Only when we know more about the behaviour of firms and households in an economy which is not so well integrated that all goods are used for the purpose they are produced for, and in which obsolescence and hence windfall capital gains and losses are widespread, can we decide whether it is possible to give precise meaning to the notion of maintaining capital intact. Until then the notion is useful only in the analysis of steady states which are not subject to technical change – which excludes most situations of interest to the economic theorist.

See Also

- ▶ Capital as a Factor of Production
- ▶ Capital Gains and Losses
- ▶ Depreciation
- ▶ Equal Rates of Profit

Bibliography

Break, G.F. 1954. Capital maintenance and the concept of income. *Journal of Political Economy* 62, February, 48–62.

Hayek, F.A. 1935. The maintenance of capital. *Economica* 2: 241–276.

Hayek, F.A. 1941a. *The pure theory of capital*. London: Routledge.

Hayek, F.A. 1941b. Maintaining capital intact: A reply. *Economica* 8: 276–280.

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

Hicks, J.R. 1942. Maintaining capital intact: A further suggestion. *Economica* 9: 174–179.

Hicks, J.R. 1958. Measurement of capital – in theory. In J.R. Hicks, *Wealth and welfare*, Oxford: Blackwell, 1981.

Hicks, J.R. 1973. *Capital and time*. Oxford: Clarendon.

Hicks, J.R. 1974. Capital controversies: Ancient and modern. Ch. 7 in J.R. Hicks, *Economic perspectives*, Oxford: Clarendon Press, 1977.

Hicks, J.R. 1979. The concept of business income. Ch. 14 in J.R. Hicks, *Classics and moderns*, Oxford: Blackwell, 1983.

Lindahl, E. 1933. The concept of income. In *Economic essays in Honour of Gustav Cassel*, ed. Gustav Cassel, 399–418. London: Allen & Unwin.

Parker, R.H., and G.C. Harcourt (eds.). 1969. *Readings in the concept and measurement of income*. Cambridge: Cambridge University Press.

Pigou, A.C. 1932. *The economics of welfare*, 4th ed. London: Macmillan.

Pigou, A.C. 1935. Net income and capital depletion. *Economic Journal* 45: 235–241.

Pigou, A.C. 1941. Maintaining capital intact. *Economica* 8: 271–275.

Pigou, A.C. 1946. *The economics of welfare*. Reprint of 4th edn. with additional Prefatory Note, London: Macmillan.

Robertson, D.H. 1933. Saving and hoarding. *Economic Journal* 43: 399–413.

Scott, M.F.G., et al. 1984. Maintaining capital intact. In *Economic theory and Hicksian themes*, ed. D.A. Collard, 59–73. Oxford: Clarendon.

Makower, Helen (1910–1998)

K. J. Lancaster

Keywords

Consumer theory; Household production; Linear methods; Makower, H

JEL Classifications

B31

Helen Makower was educated at Cambridge and obtained her doctorate from London University. From 1938 until her retirement in 1973 she taught at the London School of Economics and Political Science. In collaboration with Jacob Marschak she made a pioneering contribution to modern asset portfolio theory and to the study of labour mobility. After the Second World War her analytical insights and interest in work then being

performed at the Cowles Commission in Chicago led to her being one of the important links through which such techniques as activity analysis entered the academic scene in Britain. Her 1957 book and other papers made original contributions to the application of linear methods in economic analysis. One of her important insights was into the analogy between production and consumption, a precursor of later work on the household production and characteristics approaches to consumer theory.

Selected Works

1938. (With J. Marschak.) Assets, prices and monetary theory. *Economica* N.S. 5: 261–288.
1939–40. (With J. Marschak and H.W. Robinson.) Studies in the mobility of labour: Analysis for Great Britain. Pt. I–II. *Oxford Economic Papers* 2 (1939), 70–97; 4 (1940), 39–62.
1950. The analogy between producer and consumer equilibrium analysis. *Economica* N.S. 17: 63–68.
1957. *Activity analysis and the theory of economic equilibrium*. London: Macmillan.

Malthus and Classical Economics

S. Rashid

Thomas Robert Malthus is known to economists primarily as the author of the *Essay on . . . Population* as well as the *Principles of Political Economy*, first published in 1820. In addition, Malthus wrote some pamphlets and contributed articles to established periodicals. His first pamphlet was on the high price of provisions and contains an explicit, if rudimentary, construction of a market demand curve. Although later scholars have found much curious matter in it, the pamphlet died unnoticed. Malthus wrote some articles on monetary economics for the *Edinburgh Review* which are chiefly notable for their moderation in espousing a middle ground between the Bullionists and their opponents. His pamphlets on the Corn Laws have gained him fame as a co-discoverer of the differential fertility theory of rent. However, not only was the theory of differential rent clearly expounded by James Anderson in 1777, but also Malthus did not formulate the theory with the clarity of either David Ricardo or Edward West. Malthus's contributions to the *Quarterly Review* (1823–4) consist largely of sharper formulations of the differences between himself and the Ricardians, points which had already been raised in the *Principles*. For these reasons, my focus will be primarily upon the *Principles*. The shadow of controversy lay over almost everything Malthus wrote and Malthus's contributions will be set in sharpest focus by occasionally using Ricardo's critical comments on the first edition of Malthus's *Principles*, the *Notes on Malthus*. As both editions of Malthus's *Principles* will be quoted, the first and second editions will be referred to as *Principles* I and II respectively.

As the second, revised edition of the *Principles* was (posthumously) published in 1836, one would naturally assume that a study of the *Principles* would suffice to tell us about Malthus's principal contributions to classical economics. While the dictates of a summary article enforce such an approach, it is important to point out that several complexities of Malthus's overall impact are thereby ignored. For example, the methodology, and sometimes the substantive conclusions, of the *Essay* and the *Principles* are very different. The tendency to extremes that Malthus deplores in the *Principles*, he exemplifies in the *Essay*. Secondly, several of the positions later espoused in his pamphlets, or in the *Principles*, were originally broached in the *Essay*. One such example would be the chapter on bounties on corn in the *Essay*. Thirdly, Malthus began as a strong supporter of Agrarianism, of the superiority of Agricultural over Manufacturing wealth, and quietly modified his position so as to occupy a middle ground as time went on. Fourthly, the claim of Malthus to be the heir to Adam Smith became highly questionable after the rise to fame of David Ricardo. Finding himself on the defensive,

Malthus increasingly turned to proclaiming that he was following Smith and that Ricardo was not. Such repeated claims served to hide the fact that, on some important issues, Malthus broke sharply with the corpus of Smithian thought. Both in the *Essay* and in the second edition of the *Principles*, Malthus challenged the idea that all nations would necessarily prosper through Free Trade. He insisted on the significant qualification that the products of such countries must be complementary if mutual gains are to be assured.

> It is . . . a just and general rule in political economy, that the wealth of a particular nation is increased by the increasing wealth and prosperity of surrounding states; and that, if these states are not successful competitors in those branches of trade in which the particular nation had excelled, their increasing wealth must tend to increase the demand for its products, and call forth more effectively its resources. But if this rule be repeatedly insisted upon without noticing the above most important limitation, how is the student in political economy to account for some of the most prominent and best attested facts in the history of commerce? How is he to account for the rapid failure of the resources of Venice under the increasing wealth of Portugal and the rest of Europe, after the discovery of a passage to India by the Cape of Good Hope; the stagnation of the industry of Holland, when the surrounding nations grew sufficiently rich to undertake their own carrying trades . . . It is not favourable to the science of political economy, that the same persons who have been laying down a rule as universal should be obliged to found their explanations of most important existing phenomena on the exceptions to it. It is surely much better that such a rule should be laid down at first with its limitations (*Principles* II, 10–11).

Method

The principal difference in method between Malthus and Ricardo lay in the extent to which they believed simple models could satisfactorily illuminate reality. Both sides were agreed as to the operation of multiple causes in economic affairs but drew quite different conclusions therefrom. In the Introduction to the *Principles*, Malthus directed his strictures at the Ricardian model:

> In political economy the desire to simplify has occasioned an unwillingness to acknowledge the operation of more causes than one in the production of particular effects; and if one cause would account for a considerable portion of a certain class of phenomena, the whole has been ascribed to it without sufficient attention to the facts, which would not admit of being so solved (*Principles* II, 5).

While Ricardo fully assented to the operation of multiple causes, he drew the opposite conclusion from this fact – namely, that complex models cannot be adequately handled and, rather than build models one could get no definite conclusions from, it was more illuminating to build simple models, which, despite their simplicity, were somehow 'representative'. Malthus could not agree. To him, 'The principal cause of error and of the differences which prevail at present among scientific writers on political economy, appears to me to be a precipitate attempt to simplify and generalize' (*Principles* II, 4).

For example, the discussion of wages in the chapter on Wages in Ricardo's *Principles* is much more sophisticated than his use of an iron-law of subsistence wages in his subsequent policy analysis. Wesley Mitchell attributes this limitation of Ricardo's policy analysis to his primitive analytical apparatus, which forced him to treat one variable as a constant when dealing with a problem involving two or more variables. This limitation was reinforced by Ricardo's insistence upon definite results. In Mitchell's words, Ricardo 'was never patient with any proposition that was surrounded by a lot of "ifs" and "buts" and "to a degrees". He wanted a clear-cut view.' Malthus was well aware of this pitfall and wrote of 'a still greater disinclination to allow of modifications, limitations and exceptions to any rule or proposition' (*Principles* II, 6).

Ricardo's great partiality for single causes may be more clearly seen from his dispute with Malthus over the determinants of profits. In Ricardo's 'corn-model' the actual profit-rate is left indeterminate but it is predicted that the rate will fall over time. This prediction is based upon the necessity of having to take recourse to poorer land with the growth of population. Since more labour is required to grow a bushel of wheat on the marginal no-rent land, and since wages are fixed, profits must of necessity fall. Malthus did

not deny the truth of such an analysis but questioned its relevance. For practical problems Malthus felt that the rate of profit was effectively set by the causes which determined the final price at which a commodity would sell, i.e. by forces akin to those describing a 'brisk' market. Malthus contrasted thus the respective strengths of the two forces.

> . . . But though this [Ricardian] principle is finally of the very greatest power, yet its progress is extremely slow and gradual; and while it is proceeding with scarcely perceptible steps to its final destination, the second cause is producing effects which entirely overcome it, and often for twenty or thirty, or even 100 years together, make the rate of profits take a course absolutely different from what it ought to be according to the first cause (*Principles* II, 282).

In the *Notes on Malthus*, Ricardo sometimes appears to agree that some concessions may be in order. Despite such wavering in the *Notes*, the concluding paragraph of Ricardo's chapter on Profits in the third edition of his *Principles*, which was issued after the above controversy, begins as follows:

> Thus we again arrive at the same conclusions which we have before attempted to establish: – that in all countries, and all times, profits depend on the quantity of labour requisite to provide necessaries for the labourers on that land or with that capital which yields no rent. The effects then of accumulation will be different in different countries, and will depend chiefly on the fertility of the land (*Principles*, 126).

Not only did Ricardo's method limit him to the study of one variable at a time but also his analysis was further handicapped by his refusal to analyse other than equilibrium positions. This limitation often prevented Ricardo from appreciating the importance of the specific stimuli and the short-run dynamics that are the stuff of economic life. Malthus had argued that 'It is the natural tendency of foreign trade . . . immediately to increase the value of that part of the national revenue which consists of profits, without a proportionate diminution elsewhere' (*Principles* I, 460). To this, Ricardo replied with an example proving nothing other than that, at a new equilibrium, the supernormal profits will have been whittled away (*Notes* 418–19).

Price Theory

Malthus borrowed most of his price theory from Adam Smith and the Physiocrats. Malthus's excellence shows itself in the systematic application and extensions he made of these sources. Smith had spoken of natural price, a long-run concept, and market price, a short-run concept; typically, Smith applied the former to factor markets and the latter to commodity markets. The former was determined by the cost-of-production and the latter was determined by demand and supply. Malthus pondered over the matter and then directly asked himself why supply and demand analysis should not be used to determine natural price as well as short-run market price? Malthus had asked the right question and further thinking led him to the right answer, namely, that the cost of production was influential only as it affected the supply curve:

> . . . when we come to examine the subject more closely, we find that the cost of production itself only influences the prices of these commodities, as the payment of this cost is the necessary condition of their continued supply in proportion to the extent of the effectual demand for them (*Principles* II, 71).

It followed that there was no longer any need to have two theories of price formation – one for the short run and one for the long run: 'But if this be true, it follows that the great law of demand and supply is called into action to determine what Adam Smith calls natural prices, as well as what he calls market prices' (*Principles* II, 71).

Malthus, having extended demand and supply to explain factor prices, was able to make a very telling criticism of the theory that cost of production determines price. Since the constituents of the cost of production, i.e. rents, profits, and wages, were themselves set by the operation of demand and supply, one could hardly escape the operation of demand and supply by appealing to the cost of production.

> But if it appear generally that the ordinary cost of production only determines the usual prices of commodities, as the payment of this cost is the necessary condition of their supply; and that the component parts of this cost are themselves determined by the same causes which determine the whole, it is obvious that we cannot get rid of the

> principle of demand and supply, by referring to the cost of production. Natural and necessary prices appear to be regulated by this principle, as well as market prices; and the only difference is, that the former are regulated by the ordinary and average relation of the supply to the demand; and the latter, when they differ from the former, are determined by the extraordinary and accidental relation of the supply to the demand (*Principles* I, 78).

Ricardo, on the other hand, maintained an implicit dichotomy between goods and factor pricing as his comments on the above passage indicate:

> In this account of necessary or natural price Mr. Malthus has in substance said the same as Adam Smith has done, in all which I fully agree, but is he not inconsistent in maintaining that natural price is regulated by supply and demand (*Notes*, 52).

The major bone of contention between Ricardo and Malthus was the cost of production theory of price. Perhaps because of his emphasis upon the long-run, Ricardo believed that an explanation based on cost of production provided a deeper understanding than the superficial demand and supply arguments. This Ricardian belief prevailed as the orthodox view for much of the 19th century.

In the *Principles* Malthus emphasized the central importance of demand and supply among the analytical tools of the economist: 'It must be allowed that of all the principles of political economy, there is none which bears so large a share in the phenomena which come under its consideration as the principle of supply and demand.' As though foreseeing the subsequent controversy that was to arise, Malthus made clear that he did not wish to emphasize either demand or supply at the expense of the other: '. . . when prices are said to be determined by demand and supply, it is not meant that they are determined either by the demand alone, or by the supply alone, but by their relation to each other' (*Principles* II, 62).

To the superb exposition of price theory contained in this chapter. Ricardo's usual answer is to speak of the role of demand and supply as real but ephemeral and transient: '. . . The market price of a commodity may from an unusual demand, or from a deficiency of supply, rise above its natural prices, but this does not overturn the doctrine that the great regulator of price is cost of production' (*Notes*, 39). The subsequent pages, however, show that Ricardo was not entirely happy with this admission. A little later, when Malthus says that a fall in the cost of production will lower equilibrium price, Ricardo seizes it as an admission of the superiority of the cost-of-production theory (*Notes*, 40): as the chapter continues, Ricardo shifts the question from the determination of market price to the determination of the supply curve (*Notes*, 42–5).

In arguing the case for demand and supply Malthus says that manufactured goods, whose production at constant costs he implicitly accepts, provide the best example of commodities whose price is determined by the cost of production. Even here, however, cost comes into play only insofar as the payment of costs is 'the necessary condition of their continued supply' (*Principles* I, 84). But if costs only act through supply, then it is valid to claim that 'the great principle of supply and demand is called into action to determine . . . natural prices as well as market prices' (*Principles* I, 75). To this Ricardo replies as follows:

> The author forgets Adam Smith's definition of natural price or he would not say that demand and supply could determine its natural price. Natural price is only another name for cost of production. When any commodity sells for that price which will repay the wages for labour expended on it, will also afford rent, and profit at their then current rate, Adam Smith would say that commodity was at its natural price. Now these charges would remain the same, whether commodities were much or little demanded, whether they sold at a high or low market price (*Notes*, 46).

Ricardo is correct in claiming the existence of a dichotomy of product and factor markets as one part of the Smithian heritage. Malthus would have done well to have emphasized that he was correcting Smith and not following him. Of the merits of his emphasis upon demand *and* supply one can only say that, if accepted, it would have rendered much of the Marshallian synthesis superfluous.

The Measure of Value

A question that occupied much, perhaps most, of Malthus's later years was that of finding the best

'measure of value', i.e. a commodity which would, at all times and places, provide an accurate account, in some absolute sense, of a nation's 'richness'. The answer clearly revolves around what one means by 'richness'. Malthus decided that the amount of labour a nation could command was the truest measure of the welfare content of its GNP, the justification being that the extent of leisure a country could devote itself to was the best measure of its well-being. In the second edition of the *Principles*, Malthus argued that the capacity to demand leisure was best measured by the units of common, agricultural labour that a country could command, because (a) adequate food being everywhere a prerequisite for the existence of leisure, the relative command over the labour that thus made leisure feasible was a good indicator of real wealth; (b) the disutility incurred by common agricultural labour was believed by Malthus to be relatively uniform across time and space, and accordingly command of one unit of it always represented the same acquisition of wealth; and (c) such labour entered directly or indirectly into the production of all commodities and therefore whatever affected it would affect the value of all commodities. Thus far, it must be said, Malthus has provided a sensible approximate solution of an insoluble problem.

Malthus, however, could not leave well enough alone. He also wanted the measure of value to serve as a unit of exchange. It is an elementary proposition that any commodity with a positive price can serve as the *numéraire* in a market; if ordinary labour has a positive price, it can of course serve as *numéraire*. The several pages Malthus devotes to this trivial point provides a curious illustration of how a mind possessed of strong common-sense can become ensnared in a near-tautology.

A more fruitful outgrowth of this concern with finding the best measure of 'richness' came about in Malthus's defence of material goods as the true repository of wealth. In opposing the more modern notion of wealth as exchangeable value, Malthus provided the following arguments. (i) What an individual is willing to consider exchangeable can vary with circumstances, for example, musical talents that were developed for their own sake can be used to teach others if the possessor feels like doing so or is compelled to do so by circumstances. (ii) Large amounts of commodities are raised in agricultural districts for domestic consumption. These goods are neither raised for exchange nor ever exchanged. But these are surely wealth in the common signification of the term. (iii) Most importantly, however, non-material goods just cannot be accounted for with satisfactory precision:

> Of the quantity and quality of the material commodities here noticed it would not be difficult to make an inventory. Many household books indeed furnish one; and knowing pretty nearly the quantity and quality of such articles a fair approximation to their value might be attained by estimating them according to the market prices of the district at the time. But in regard to immaterial objects, the difficulty seems to be insurmountable. Where is an inventory to be found, or how is one to be made of the quantity and quality of that large mass of knowledge of the individual possessors and their friends? Or supposing it were possible to form such an inventory, how could we make any moderate approaches towards a valuation of the articles it contained? (*Principles* II, 27).

It is important to note that Malthus approached the question of defining wealth solely with (somewhat narrow) accounting considerations in mind. He was so purist as to insist that only market values were acceptable and explicitly rejected the modern practice of imputing values to non-marketable outputs by using the costs of producing such outputs. If two students paid the same fees for attending school, could we say that they had both gained equally? Malthus even rejected the salaries of government clerks from the national income. While such clerks often performed exactly the same duties as a clerk in a commercial firm, only in the latter case could we be sure that the clerk was a necessary employee because he served a profit-maximizing firm.

In the process of defending the notion that material goods should be the sole components of the national income, Malthus is driven to pointing out all the difficulties national-income accountants would have to face if they adopted the broader (and more modern) notion of exchangeable value. Even though Malthus himself never estimated the national income, his careful study of the issues involved entitle him to be considered the first methodologist of national income accounting.

Say's Law

Malthus's objection to the idea that supply creates its own demand arise from his claim that, in the short run, the desires of mankind are easily satiated and may be practically considered as fixed. Since the desire to consume is thus quite limited in the short run, if the power of producing commodities be steadily increasing, then the consumers will become satiated and refuse to purchase further goods. The drop in sales resulting from this refusal to purchase causes a fall of profits and leads to a cumulative downward movement and a general stagnation of trade. Malthus's basic point is that while the increase in the supply of commodities is an ongoing, self-propelling process, since there is no similar mechanism for 'updating' the desires of consumers, the prospect for profitable sales of commodities are bleak. Yet another way of stating the argument is to say that technological progress in supply guarantees an increasing amount of commodities but wherefrom comes the technological progress in demand that would ensure a market for such produce?

The model used by Malthus to argue for the plausibility of general gluts postulates an economy consisting of only three classes, farmers, capitalists and workers. For the workings of the model, Malthus assumes that workers throughout earn a subsistence income. The remainder of what is produced is exchanged between farmers and capitalists, who thus form complementary markets for each other. Now, says Malthus, if both farmer and capitalist simultaneously decide to curtail consumption and to accumulate then the market for final products will rapidly die off. Goods become unsaleable, profits drop and general glut of commodities ensues. It is true that the desire to accumulate will initially generate a demand for investment. But investment is the auxiliary of consumption and is, in the final analysis, undertaken only to provide consumption. If the final products look unsaleable, then why should one invest at all? The entire argument is complete and self-contained; there is no need to appeal to monetary factors.

The essence of Malthus's model is indicated by the quote below:

> It is undoubtedly possible by parsimony to devote at once a much larger share than usual of the produce of any country to the maintenance of productive labour; and suppose this to be done, it is quite true that the labourers so employed are consumers as well as those engaged in personal services, and that as far as the labourers are concerned, there would be no diminution of consumption or demand. But it has already been shown that the consumption and demand occasioned by the workmen employed in productive labour can never *alone* furnish a motive to the accumulation and employment of capital; and with regard to the capitalists themselves, together with the landlords and other rich persons, they have, by the supposition, agreed to be parsimonious, and by depriving themselves of their usual conveniences and luxuries to save from their revenue and add to their capital. Under these circumstances, it is impossible that the increased quantity of commodities, obtained by the increased number of productive labourers, should find purchasers, without such a fall of price as would probably sink their value below that of the outlay, or, at least, so reduce profits as very greatly to diminish both the power and the will to save (*Principles* II, 314–15).

The greatly increased productive powers of the post-Napoleonic era placed Britain somewhat in the position of the economy described above. That people would eventually come to desire more goods Malthus had no doubt; but he was equally convinced that such desires could not be taken for granted:

> That an efficient taste for luxuries and conveniences, that is, such a taste as will properly stimulate industry, instead of being ready to appear at the moment it is required, is a plant of slow growth, the history of human society sufficiently shows; and that it is a most important error to take for granted, that mankind will produce and consume all that they have the power to produce and consume, and will never prefer indolence to the rewards of industry, will sufficiently appear from a slight review of some of the nations with which we are acquainted (*Principles* II, 321).

Since the critical reason for a general glut was the satiation of wants in the short run, the remedies suggested by Malthus naturally described ways of raising the short-run marginal propensity to consume. One way of achieving this end was by redistributing income towards the working classes, since the poor had much less trouble spending their money than the rich.

> Thirty or forty proprietors, with incomes answering to between one thousand and five thousand a year, would create a much more effectual demand for the necessaries, conveniences and luxuries of life, than a single proprietor possessing a hundred thousand a year (*Principles* II, 374).

Political considerations however prevented Malthus from seriously advocating systematic redistribution as a cure for economic depressions. Another possible remedy, and one much more amenable to Malthus's partialities, was the existence of a large group of 'unproductive consumers', i.e., individuals who had incomes but did not produce market-oriented outputs. Clergymen formed a good example of a group who contributed to the aggregate demand without adding to the market supply. It is worth noting that the role of defence expenditures as a means of maintaining aggregate demand is a direct application of this argument, a point frequently missed by radicals who have followed Karl Marx in refusing to see substantial merit in Malthus's economics.

Malthus felt unable to state any determinate proportion between productive and unproductive consumers required for the maintenance of aggregate demand. Ricardo would have none of this. Perhaps because he failed to grasp Malthus's specific definition of the term, Ricardo believed unproductive consumers to be as useful as fires and the determination of their exact proportion was no puzzle.

> I should find no difficulty to determine. They may be useful for other purposes but not in any degree for the production of wealth (*Notes*, 422).

A final remedy suggested by Malthus for gluts, and one reflective of his emphasis upon short-run dynamics, was the creation of new wants. One of the great benefits of finding new markets, according to Malthus, was that new commodities would be introduced, which in turn would stimulate consumers desires and make them more eager to spend. Referring to the depression of the cotton industry of Glasgow, Malthus said,

> It is specifically to overcome the want of eagerness to purchase domestic commodities that the merchant exchanges them [with foreigners] for others more in request. Could we but so alter the wants and tastes of the people of Glasgow as to make them estimate as highly the profusion of cotton goods which they produce, as any articles which they could receive in return for them under a prosperous trade, we should hear no more of their distresses (*Principles* II, 392–3).

It would be a mistake to consider the satiation of wants argument as peculiar to Malthus. Many of his contemporaries made the same point, although perhaps with less forcefulness. It would even appear to be the main objection to an acceptance of Say's Law by many political economists of the period 1820–1840. The prevalence of such objections was perhaps responsible for the clarification of the insatiability argument in this period. Both Nassau Senior and W.F. Lloyd spelled out at length the fact that satiation for each good was reasonable; it was only the continuous creation of new goods that made the insatiability argument acceptable. Malthus was aware of this point and had noted that the creation of such desirable new goods was precisely the root of the difficulty:

> It will be readily allowed that a new commodity thrown into the market, which, in proportion to the labour employed upon it, is of higher exchangeable value than usual, is precisely calculated to increase demand . . . But to fabricate or procure commodities of this kind is the grand difficulty; and they certainly do not naturally and necessarily follow an accumulation of capital and increase of commodities (*Principles* I, 356).

The belief that Malthus attributed gluts to a shortage of aggregate income is a later and incorrect construction. Supporters of Say's Law had pointed out that since everything that was produced was income to somebody the aggregate of rent, wages and profits must suffice to buy aggregate output. Malthus fully accepted the point that the income was there; what he questioned was whether people would spend their incomes.

> It has been repeatedly conceded, that the productive classes have the power of consuming all that they produce; and, if this power were adequately exercised, there might be no occasion, with a view to wealth, for unproductive consumers. But it is found by experience that, though there may be the power, there is not the will; and it is to supply this will that a body of unproductive consumers is necessary (*Principles* I, 489).

Whatever the workers may desire to consume, they certainly did not possess the ability to

purchase the entire output; the landlords and unproductive consumers were not numerous enough to keep demand high while the capitalists produced not to consume but to gain prestige and save money for their children. Since the last class did most of the producing, their niggardly wants contributed to the rise of a general glut.

The persistence of Say's Law as one of the dogmas of classical economics is largely due to the unwillingness of Ricardo and Say to follow through the consequences of the concessions they were forced to make to Malthus. When asked where the increased productions were to find consumers, Ricardo replied:

> If they were suited to the wants of those who would have the power to purchase them, they could not fail to find purchasers, and that without any fall of price (*Notes*, 303–304).

John Chipman (1965, p. 3) has aptly noted the greatly reduced significance of this version of Say's Law: '[Ricardo's] version of the classical principle takes much of the punch out of Say's Law; supply creates its own demand, yes – provided, of course, that not more is supplied than will be demanded.' J.B. Say made a similar concession and this makes the continuance of 'Say's Law' in the published works of both Ricardo and Say an unfortunate anomaly.

In more recent times, a failure to understand Malthus has been largely fostered by the eulogy of Lord Keynes. By stating that 'Malthus is dealing with the monetary economy in which we happen to live', Keynes suggested that Malthus's analysis was somehow concerned with monetary factors, which is not the case at all, and the search for monetary influences has misled some later scholars. If we must use Keynesian concepts, then Malthus's argument was that the Consumption Function became horizontal (in the short run) at levels of income greatly in excess of those consumers are habituated to. As producers myopically extrapolated an increasing consumption function, expectations were bound to be dashed if the economy were sufficiently productive. The tendency to ask whether Malthus was an early Keynesian has also been fruitful in generating mistakes regarding Malthus's thoughts. Since the satiability of wants plays no role in Keynesian analysis it is easy to answer such questions in the negative, but it is essential to note that such questions miss the real point. Malthus tried to analyse the British economy between 1810 and 1820 and his analysis should be judged primarily in terms of the problem he set out to solve.

Conclusion

Thomas Robert Malthus made several notable contributions to economic theory. He provided the first clearly reasoned arguments for basing National Income accounts solely on material goods; in the process he discovered most of the problems that later generations of National Income Statisticians have had to solve. He deserves some credit for re-discovering the differential fertility theory of rent. Malthus greatly extended the scope of demand and supply as fundamental economic concepts. In particular, he provided a treatment of factor markets in these terms, thereby advancing economic theory well beyond the stage at which Adam Smith had left things. He was equally inventive in observing some major limitations to the body of Smithian analysis. The most famous such exception is the possibility of a general glut of commodities whenever accumulation becomes overly rapid; equally challenging was Malthus's statement that free International Trade was a boon only to countries which produced complementary and not competitive goods. Apart from these specific contributions, his balanced and judicious methodological guidelines provide useful reading even today. Even after making allowance for exaggeration, the basic truth of Lord Keynes's conclusion appears sound: 'if only Malthus, instead of Ricardo, had been the parent stem from which 19th-century economics proceeded, what a much wiser and richer place the world would be today.'

See Also

- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Malthus's Theory of Population
- ▶ Ricardo, David (1772–1823)

Bibliography

Chipman, J. 1965. A survey of the theory of international trade: Part 1. *Econometrica* 33(3): 477–519.
Malthus, T.R. 1820. *Principles of political economy considered with a view to their practical application*. London: Murray. 2nd ed, London: Pickering, 1836.
Mitchell, W. 1967. In *Types of economic theory*, ed. J. Dorfman. New York: Kelley.
Ricardo, D. 1817. On the principles of political economy and taxation. In *The works and correspondence of David Ricardo*, vol. I, ed. P. Sraffa. Cambridge: Cambridge University Press. 1951.
Ricardo, D. 1951. Notes on Malthus. In *The works and correspondence of David Ricardo*, vol. II, ed. P. Sraffa. Cambridge: Cambridge University Press.

Malthus, Thomas Robert (1766–1834)

J.M. Pullen

Abstract

A brief biographical sketch precedes a discussion of important methodological features of Malthus's work, especially his 'doctrine of proportions' and the need for moderation and balance in economic principles and policies. His principle of laissez-faire admitted exceptions; and, although his principle of population warned of over-population, he acknowledged the potential advantages of population growth. His ideas on the Poor Laws, the Corn Laws, Say's Law, and the relation between saving and investment are discussed; and the roles given to effective demand and to distribution as a factor of production, especially the distribution of landed property, are emphasized.

Keywords

Classical economics; Cobbett, W.; Condorcet, M. de; Corn Laws; Distribution theories: classical; Doctrine of proportions; Economic growth; Effective demand; Free trade; Godwin, W.; Happiness; Hoarding; Keynes, J. M.; Laissez-faire; Living standards; Malthus, T. R.; Malthus's theory of population; Marx, K. H.; Mill, J.; Old Poor Law; Optimum; Perfectibilism; Population growth; Protectionism; Redistribution of wealth; Saving–investment equality; Say, J.-B.; Say's Law; Unproductive consumption

JEL Classifications

B31

Malthus has the unusual distinction not only of being a founder of classical economics – mainly because of his principle of population – but also of being instrumental in attempts to overthrow classical economics, mainly because of his principle of effective demand and its influence on John Maynard Keynes.

The most comprehensive and authoritative source of biographical information on Malthus is James (1979), from which the following brief details have been largely derived. Additional information can be found in the first edition of *The New Palgrave: A Dictionary of Economics* (Pullen 1987), Malthus (1989b, pp. xv–lxix), and the *Oxford Dictionary of National Biography* (Pullen 2004). Malthus was born on 13 February 1766, near Wotton, in the county of Surrey, England, and died on 29 December 1834, at Bath. He was buried in Bath Abbey where there is a commemorative plaque. Although he was baptized Thomas Robert, he used his full name only in formal situations; in less formal correspondence he signed himself T. Robert Malthus or Robert Malthus, and was known to family and close friends as Robert or Bob.

He was the son of Henrietta (née Graham) (1733–1800) and Daniel Malthus (1730–1800). The latter, having inherited independent means, cultivated literary, artistic, scientific and theatrical interests. He was an admirer and correspondent of Rousseau, who once visited the family home soon after Malthus's birth. The extensive library of Daniel Malthus was eventually passed on to Malthus and, supplemented by acquisitions of his own and other family members, is now held in Jesus College, Cambridge.

Malthus graduated in 1788, and in 1789 was ordained deacon with title to a stipendiary curacy at the small chapel at Okewood in the parish of Wotton. He was ordained priest in 1791, was appointed non-resident Rector of Walesby in Lincolnshire in 1803, and succeeded to the perpetual curacy of Okewood in 1824. He married in 1804 and had three children, but no grandchildren. In 1805 he was appointed to the East India College as 'Professor of General History, Politics, Commerce and Finance', a title later altered to 'Professor of History and Political Economy'. He held the post for the rest of his life, residing in the College at Haileybury, near Hertford. As well as performing his teaching duties, he preached regularly in the college chapel. The important collection of Malthus manuscripts held at Kanto Gakuen University in Japan (Malthus 1997, 2004) contains four of his sermons. They corroborate the statement of his colleague William Empson: 'Mr. Malthus was a clergyman – a most conscientious one, pure and pious. We never knew one of this description so entirely free of the vices of his caste' (Empson 1837, p. 481). His main publications were *An Essay on the Principle of Population,* first published in 1798, with five further editions in 1803, 1806, 1807, 1817 and 1826, and *Principles of Political Economy,* first published in 1820 with a posthumous second edition in 1836. He also published at least 20 smaller works – his authorship of a 21st is disputed – and evidence he gave at two public enquiries can be found in the published reports. There is a full list of his publications in *The New Palgrave* (Pullen 1987) or in Malthus (1986, vol. 1, pp. 41–4). He engaged in extensive correspondence throughout his career, with Ricardo and many others. More than 230 letters to and from over 50 correspondents are known to have survived.

Malthus's Methodology: 'The Doctrine of Proportions'

Before considering particular aspects of Malthus's political economy, it is important to understand some of the peculiar features of his methodology. Failure to do so has resulted in many misunderstandings and unnecessary disagreements among commentators.

One of the most important, but one of the most unrecognized, aspects of Malthus's methodology was the principle that he called the 'doctrine of proportions'. This was the traditional ethical notion of the just mean or middle way. As Leslie Stephen (1893) said, in the first *Dictionary of National Biography,* Malthus was always 'a lover of the golden mean'. The distinctive innovation of Malthus lay in applying the concept to political economy, and in giving it such a prominent and consistent role.

He stated that his aim was to show 'how frequently the doctrine of proportions meets us at every turn, and how much the wealth of nations depends upon the relation of parts'. It was his view that 'all the great results in political economy, respecting wealth, depend upon *proportions'*, and warned that the 'tendency to extremes is one of the great sources of error in political economy, where so much depends upon proportions'. He added that 'It is not, however, in political economy alone that so much depends upon proportions, but throughout the whole range of nature and art' (1989b, vol. 1, pp. 352, 432; vol. 2, pp. 252, 269, 278).

Malthus's doctrine of proportions is thus essentially the same as the concept of the optimum. Although he did not use the term 'optimum', he must be recognized as having been one of the first to introduce the concept of the optimum into economics. In giving this central role to the doctrine of proportions, he has in effect said that *the* economic problem is the problem of balance, not the problem of choice.

But, despite his widespread use of the doctrine of proportions, Malthus recognized that precise determination of optimum points would be difficult. In discussing the optimum level of saving, he acknowledged that 'the resources of political economy may not be able to ascertain it' (1989b, vol. 1, p. 9), and, in discussing the just means for saving and the division of landed property, he said 'the extremes are obvious and striking, but the most advantageous mean cannot be marked' (1989b, vol. 1, p. 10).

The moderation and balance implied by the doctrine of proportions was evident in Malthus's

personal temperament. Bishop Otter, who knew Malthus for nearly 50 years, said that he 'scarcely ever saw him ruffled, never angry, never above measure, elated or depressed', and that Malthus possessed 'a degree of temperance and prudence, very rare at that period, and carried by him even into his academical pursuits' (in Otter 1836, pp. xxxii, xlix); and William Empson said in reference to the doctrine of proportions: 'The lesson which he sought to impress on others, he faithfully applied to himself; and so successfully, that few characters have ever existed of more perfect symmetry and order' (Empson 1837, pp. 476–7).

Malthus has been given credit for introducing or propagating, either alone or with others, a number of key ideas in the history of economics; notably, the principle of population, the law of diminishing returns, and the role of effective demand. The doctrine of proportions could be added to the list.

Limitations and Exceptions

Another facet of Malthus's methodology was his insistence on limitations and exceptions to the general principles in political economy. This could be seen either as a corollary of his doctrine of proportions or as another way of expressing the same doctrine. He believed that there are some general principles in political economy to which exceptions are 'most rare', but added 'yet there is no truth of which I feel a stronger conviction than that there are many important propositions in political economy which absolutely require limitations and exceptions' (1989b, vol. 1, p. 8). In this respect, he departed from the absolutist and universalist aspirations of some of his contemporaries, who, anxious to promote the scientific credentials of political economy, pretentiously declared them to be 'laws'. He was critical of the 'precipitate attempt to simplify and generalize', which he regarded as the 'principal cause of error, and of the differences which prevail at present among the scientific writers on political economy' (1989b, vol. 1, pp. 5–6).

Malthus has been accused, in his own day and now, of lacking in logic, especially by comparison with Ricardo. The accusation that his views did not constitute a logical and coherent system appears to have emanated from a failure to appreciate that his views were formulated in the context of his doctrine of proportions, and that he believed exceptions and limitations frequently have to be admitted when principles are used to formulate policies for application to particular real-world circumstances.

Laissez-Faire and Government Intervention

Malthus strongly supported the principle of laissez-faire or freedom of trade: 'the wealth of nations is best secured by allowing every person, as long as he adheres to the rules of justice, to pursue his own interest in his own way', and 'governments should not interfere in the direction of capital and industry, but leave every person, so long as he obeys the laws of justice, to pursue his own interest in his own way'. He described this as a 'great principle' and as 'one of the most general rules of political economy' (1989b, vol. 1, pp. 3, 13, 518).

But he also argued that some exceptions to the principle of laissez-faire have to be recognized, and that the principle of non-interference is 'necessarily limited in practice' (1989b, vol. 1, 18–19, 525). He believed that there are certain duties that belong to the government – for example, in areas such as education; support of the poor; construction and maintenance of roads, canals, and public docks; colonization and emigration; and the support of forts and establishments in foreign countries – although he recognized that there may be differences of opinion about the extent to which government should share in such matters. In particular, the 'necessity of taxation ... impels the government to action, and puts an end to the possibility of letting things alone' (1989b, vol. 1, pp. 18–19).

Thus, although Malthus strongly supported the principle of laissez-faire, his support, like that of Adam Smith, was pragmatic and conditional rather than dogmatic and absolute. There was, however, a major difference in their conception

of the laissez-faire principle. In what Donald Winch has described as 'an attack on a central feature of the *Wealth of Nations*' and as 'a major qualification to Smith's system of natural liberty', Malthus doubted whether economic growth has always been, or will always be, advantageous to the mass of society. Malthus criticized Smith's view that the economic growth of Britain during the 18th century had improved the living standards of the labouring classes; he recognized that investments in trade and manufacturing had benefited individual capitalists, but argued that they were of less benefit to society as a whole. Thus Malthus raised the possibility of conflict between economic growth and human happiness, and implied that interventionist welfare policies by government might be justified. In this respect, as Winch has argued, 'if general allegiance to the system of natural liberty, as interpreted by Smith and upheld under the different circumstances by some of his followers, is the hallmark of an orthodox political economist during the first half of the nineteenth century, Malthus occupies a decidedly ambivalent position' (Winch 1987, pp. 32, 59–61, 76–7).

Population

Malthus's first published work – *An Essay on the Principle of Population* (1798) – was written primarily to controvert the perfectibilist notions of Godwin and Condorcet. He believed that the growth of population presented a major obstacle to unlimited human progress. He argued that population will constantly tend to exceed the food supply, with the result that human progress will be neither rapid nor unlimited, and will be accompanied by sufferings and evils arising from the operation of unavoidable checks to population growth.

To support his views on the threat of overpopulation to human progress, Malthus introduced the notion of the two ratios. He argued that population will tend to increase in a geometrical ratio (1, 2, 4, 8 ...) doubling every 25 years; but the food supply will increase only in an arithmetical ratio (1, 2, 3, 4 ...). He believed that the population of 'this Island' was then about seven million, and that after the first 25 years it would reach 14 million, after 50 years 28 million, after 75 years 56 million, and so on. But the utmost that could be expected for the supply of food is that there would be sufficient to feed 14 million after 25 years, 21 million after 50 years, 28 million after 75 years, and so on. Thus, after the first 25 years, the food supply would become insufficient, and any further progress in the size of the population and the standard of living would be impossible. He concluded that this argument is conclusive against the perfectibility of the mass of mankind.

Opinions differ on whether Malthus's principle of population depends essentially on the empirical accuracy of these ratios or whether they were intended merely as approximate tendencies, or as a mathematical metaphor. Whatever his intention, there is no doubt that the ratios have exerted a powerful rhetorical influence in promoting his message and his fame.

Malthus was not the first writer to issue a warning about the dangers of overpopulation, as he himself acknowledged, but for a variety of reasons his arguments have become the best-known, and have exerted a great influence on human thought and human affairs. He alerted the world to the problem of overpopulation, and his views continue to affect the population policies of governments through the world today.

Having presented his basic arguments in the first two chapters of the *Essay,* Malthus then proceeded to discuss the 'checks' to population. As he said in his first postulate, people cannot live without food, and therefore it would be impossible to have a situation where 28 million people were in existence but the food supply was adequate for only 21 million. There must therefore be some mechanisms or checks whereby populations are prevented from exceeding the food supply. The bulk of the *Essay*, especially in the much enlarged later editions, was devoted to a detailed description of the checks that have operated in different countries and at different times.

He classified the checks as either positive checks that reduce normal life expectancy and increase the death rate, or preventive checks that

reduce the birth rate. Among his list of positive checks he included common diseases, epidemics, wars, plagues, pestilence, famines, infanticide, unwholesome occupations and habitations, severe labour, exposure to the seasons, extreme poverty, bad nursing of children, great cities, and excesses of all kinds.

The preventive checks, described in circumspect language ('vicious customs with respect to women'), included prostitution and birth control, but the only preventive check that he approved of and advocated was prudential restraint, by which he meant delaying marriage until sufficient resources of food, accommodation and other necessaries are available to provide the parents and the expected number of children with an acceptable standard of living. He noted that prudential restraint is practised, and should be practised, by those who want to maintain after marriage the social and economic status they enjoyed before marriage. The case for prudential restraint was even more vigorously argued in the later editions, where those who marry and raise children without ensuring that they have sufficient resources are accused of irresponsible and immoral behaviour.

He also classified the population checks as either vice or misery, but did not clearly show how the vice-and-misery classification is related to the positive-and-preventive. Presumably he meant that, among the positive checks, some, such as war and infanticide, are vices, and all lead to misery; and among the preventive checks all except prudential restraint are vices, and all are likely to lead to misery; and although prudential restraint is a virtue, not a vice, it often leads to vice.

> Restraints upon marriage are but too conspicuous in the consequent vices that are produced in almost every part of the world; vices, that are continually involving both sexes in 'inextricable unhappiness'. (1986, vol. 1, p. 28)

In admitting that prudential restraint might also be a cause of misery, he might have been speaking from personal experience, being in 1798 a 32-year-old bachelor with an income as a curate insufficient to support a wife and family in a socially acceptable manner.

In the second and later editions, he introduced the expression 'moral restraint', by which he meant prudential restraint conducted in accordance with Christian moral precepts regarding premarital sex, but the *concept* of moral restraint is *implicit* in the first edition. It is unlikely that, in advocating prudential restraint, Malthus as a Protestant clergyman would have intended to condone prudential restraint that was accompanied by immoral sexual behaviour.

In the second (and later) editions of the *Essay,* he softened some of the harshest conclusions of the first by arguing that, if people could be made aware of the harm done by improvident procreation, then moral restraint, though still a difficult challenge, could be practised without causing misery and without leading to vicious practices. He objected to contraception on moral grounds and also because, by facilitating control of the birth rate and reducing the pressure of population, it would remove one of the incentives needed to overcome our natural indolence, to promote economic growth, and to encourage the 'growth of mind'. It is ironical that the expression 'Malthusian practices' became synonymous with contraception, and that contraception has become the method most commonly adopted throughout the world to control population. The world has responded to Malthus's warnings of the danger of overpopulation by adopting a remedy he strongly rejected.

Arguments in Favour of Population Growth

The popular and superficial view of Malthus is that he was opposed to population growth. But there are numerous instances in his writings which show that he regarded an increase of population, under certain conditions, as desirable in itself, and as a necessary cause of economic growth. For example, he spoke of the 'pursuit of the desirable object of population' (1986, vol. 3, p. 455); and, referring to the possibility of a great increase of population in Ireland in the 19th century, he said 'so great an increase of human beings, if they could be well supported, would be highly

desirable' (1986, vol. 4, p. 32). In a similar vein he said: 'That an increase of population, when it follows in its natural order, is ... a great positive good in itself, ... I should be the last to deny' (1989a, vol. 1, p. 439). And those who use Malthus to support a policy of population reduction forget that on one occasion he argued that a diminution of population would be harmful: 'It is evidently therefore regulation and direction which are required with regard to the principle of population, not diminution or alteration' (1989a, vol. 2, p. 94).

Some of his most forceful statements in favour of population growth occurred in the appendices added to the third (1806) and fifth (1817) editions of the *Essay,* in response to critics who had accused him of being anti-population. The fact that these appendices have been omitted from some modern reprints of the *Essay* might explain the limited awareness of his pro-population ideas.

Malthus's pro-population views can even be seen when he was advocating prudential restraint: 'Prudential habits with regard to marriage carried to a considerable extent, among the labouring classes of a country mainly depending upon manufactures and commerce, might injure it' (1989b, vol. 1, p. 236; vol. 2, p. 215). This is a surprising argument, given that he had said that the preventive check of prudential restraint should be a principal remedy for overpopulation. It shows that he wished the doctrine of proportions to be applied as a check to the preventive check!

Malthus's pro-population views can also be seen in his statements on population as a *necessary* cause of economic growth. He admitted that population growth alone will not promote economic growth; for example, he argued that 'the increase of population alone ... does not furnish an effective stimulus to the continued increase of wealth' (1989b, vol. 1, pp. 347–8), 'population alone cannot create an effective demand for wealth' (1989b, vol. 1, p. 350) and 'encouragements to population ... will not alone furnish an adequate stimulus to the increase of wealth' (1989b, vol. 1, p. 351). But he also stated:

> That a permanent increase of population is a powerful and necessary element of increasing demand, will be most readily allowed. (1989b, vol. 1, p. 347; in the second edition, 'permanent' was changed to 'continued')
>
> and
>
> That an increase of population, when it follows in its natural order, is ... absolutely necessary to a further increase in the annual produce of the land and labour of any country, I should be the last to deny. (1989a, vol. 1, p. 439)

In other words, although Malthus recognized that population growth is not a *sufficient* cause of economic growth, he nevertheless regarded it as a *necessary* cause.

In some circumstances, according to Malthus, an increase in population will bring about a decrease in living standards; but in other circumstances it will bring about an increase in living standards, and a decrease in population will bring about a decrease in living standards. Living standards can be both a direct and an inverse function of population. Some critics would regard this as self-contradictory, as proof of his lack of logic, and as a justification for William Cobbett's epithet 'muddle-headed Malthus'. Others would see it as a reasonable, parabolic application of the doctrine of proportions.

Theological Aspects of the Principle of Population

The early chapters of the first edition of the *Essay* have a rather pessimistic tone. They appear to be saying that the pressure of population against the food supply will keep the mass of the population at or near subsistence level, and that this struggle between food and population will be accompanied by miseries and vices. However, in the last two chapters of the first edition of the *Essay* Malthus explored the theological implications of his principle of population. His published contributions in theology are too limited for him to be considered as a theologian in a professional sense, but his theological views are interesting in their own right, because of their heterodox nature, and because they seem to have been presented, not as a mere afterthought or pious homily, but in an attempt to integrate his principle of population into a comprehensive world view, in opposition to that of Godwin and Condorcet. As a Christian

minister he would have been concerned to show that his view of population did not conflict with Christian ideas about the nature of God. He had due cause for concern. In saying that misery and vice can come from obeying the biblical injunction to go forth and multiply, he was accused by some critics of blasphemously denying the Creator's omnipotence, omniscience and benevolence.

However, in the last two chapters of the first edition of the *Essay* he argued, on the contrary, that population pressure is providentially ordained by God as a means whereby human development ('the growth of mind') is stimulated. He argued that the constant pressure of population against food supply, although it might produce some moral and physical evils, would also produce an overbalance of good. The first edition of the *Essay* thus finished on a note of moral and theological optimism.

The last two chapters were omitted from subsequent editions of the *Essay*. Comments contained in his correspondence (1997, pp. 73–7) and remarks from other contemporaries indicate that the omission occurred at the instigation of friends. Some commentators interpret the omission as a recantation. Others find traces of his theology in the later editions, and argue that his growth-of-mind theology remained an essential, if only implicit, framework throughout all editions of the *Essay*. They argue that to ignore its theological aspect is to ignore an essential element of his total population theory and, contrary to his intentions, to reduce the *Essay* to a mere economic or political tract.

Poor Laws

Although Malthus believed that population pressure was a phenomenon common to most societies, he argued that the problem had been exacerbated in England by the Poor Laws. They were intended to alleviate poverty, but only succeeded in creating the poor they sought to maintain. They encouraged people to marry too early and have large families, in the expectation that food and accommodation would be provided for them; and they discouraged hard work and the development of productive skills.

In his earlier writings Malthus had argued for abolition of the Poor Laws, both as a principle and as a practical policy; but in later writings and in correspondence (see James 1979, p. 450; Winch 1996, pp. 320–1) his position moved from complete abolition to gradual abolition, and then to administrative reform, arguing that a fundamental change involving complete abolition would present practical and political difficulties, and that the most that could be achieved in the current circumstances would be an amelioration of the present system through improved administration.

This is an example of his insistence on the need for limitations and exceptions in the practical application of general principles. He did not see any contradiction in subscribing to the idea in principle while at the same time rejecting it as a practical policy for a particular place and time. Another example of this feature of his methodology occurred in his views on the Corn Laws.

Corn Laws

Although Malthus strongly supported the principle of laissez-faire, he published a pamphlet in 1815 supporting the retention of the Corn Laws which prohibited for example, the import of wheat when the home price fell below 80 shillings a quarter. This radical departure from laissez-faire caused dismay among other political economists and among his Whig friends who opposed the protectionist policy of the Tory government. In admitting this exception to the principle of laissez-faire, he was in effect reaffirming his view that, unlike the laws of mathematics, the principles of political economy should not be applied in an absolutist and universalist manner.

It has been argued that in his later years Malthus changed his mind and recanted his earlier support for the Corn Laws. The arguments for and against this change-of-mind hypothesis have been elaborated elsewhere (Hollander 1992, 1995; Pullen 1995), and are too detailed to be repeated here. It would probably be fair to say that, on the basis of the textual and contextual evidence so far presented, there is no clear, unambiguous statement of a recantation by Malthus.

But it should also be said that Malthus was strongly in favour of the principle of free trade, and that he strongly regretted the need for an exception in the case of the Corn Laws. It is obvious from his writings and correspondence that, if the circumstances that necessitated the exception were removed, he would have gladly removed his support for agricultural protection.

Economic Growth, Effective Demand and Say's Law

Malthus's views on economic growth are to be found scattered throughout his many publications, with his most systematic (but not comprehensive) treatment of this topic in the final chapter of the *Principles,* namely. chapter 7, 'On the Immediate Causes of the Progress of Wealth'. He divided the immediate causes of progress into two categories: 'the powers of production' and 'the means of distribution'. On the production side he discussed four causes: population, accumulation, soil fertility and inventions (which, by combining the second and the fourth, could be reclassified as labour, capital and land). On the distribution side he discussed three causes: the division of landed property, commerce (internal and external), and unproductive consumers. His views on the production side were unremarkable at the time, and would be quite acceptable in standard texts today, but his views on the distribution side have proved to be controversial because of their emphasis on the role of effective demand.

By effective or effectual demand Malthus meant the *power* to purchase at a price sufficient to cover the vendor's costs and required profit, combined with the *willingness* to purchase. His distinction between power and will, or means and motives, was a recurring theme in his political economy. He stressed that production requires more than the power to produce; it requires also the motive to produce, which comes from effective demand.

> ... the powers of production, to whatever extent they may exist, are not alone sufficient to secure the creation of a proportionate degree of wealth. Something else seems to be necessary in order to call these powers fully into action. This is an effectual and unchecked demand for all that is produced. (1989b, vol. 1, p. 413; vol. 2, pp. 263, 447)

The powers of production will be 'called into action, in proportion to the effective demand for them' and 'General wealth, like particular portions of it will always follow effective demand (1989b, vol. 1, pp. 414, 417). In effect he was saying that demand-side forces are as powerful and as necessary as the supply-side forces of natural resources, capital accumulation, division of labour, and so on.

He believed that an important cause of an adequate level of effective demand was the existence of a body of 'unproductive consumers', who purchase material products but do not produce material products. They would include menial servants, military personnel, actors, clergymen and other service providers. The concept was completely misunderstood by Ricardo who said that unproductive consumption is as useful as a fire in a warehouse or the destruction of war. Malthus later recognized that the term 'unproductive' had pejorative implications, and altered it to 'the provision of services'. However, the concept could also include those who live on their investments in the national debt and those whose wealth enables them to consume without either producing material goods or providing services, thus inviting Marx's description of Malthus as a protector of the ruling classes and the idle rich.

Malthus's views on effective demand were largely rejected during his lifetime, and largely ignored for the next hundred years. It was generally believed with James Mill, Jean-Baptiste Say and others that the purchasing power generated during the production process would be sufficient for all the products to be sold, that aggregate demand deficiency would never be a cause of economic decline, and that a general glut of products would be impossible. This view, known as Say's Law or Mill's Principle, and popularly expressed as 'supply creates its own demand', became a standard theme of classical economics, and still finds its supporters, even though Malthus showed, and Say virtually admitted, that its validity relies on a tautological definition of 'supply' and 'product'. The experience of the depression of the 1930s

and the publication of J.M. Keynes's *General Theory* (1936) cast doubt on this conventional wisdom of Say's Law, and rescued Malthus's views on effective demand from oblivion.

Effective Demand and the Division of Landed Property

Malthus's views on effective demand as a stimulus to economic progress led him to advocate a wider distribution of wealth, because 'Practically it has always been found that the excessive wealth of the few is in no respect equivalent, with regard to effective demand, to the mere moderate wealth of the many' (1989b, vol. 1, p. 431). But this redistribution of property has often been neglected, and sometimes even denied, in the secondary literature. Karl Marx, in particular, misinterpreted Malthus in this regard, and some other commentators appear to have taken their views of Malthus from Marx; and, like Marx, have not bothered to test them against Malthus's text.

Admittedly, there are passages in some parts of Malthus's writings that support a pro-landlordism interpretation. But in other passages he was critical of the distribution of land and other property, and described the existing maldistribution as unjust and as an impediment to economic growth; for example.

> A very large proprietor, surrounded by very poor peasants, presents a distribution of property most unfavourable to effective demand ... Thirty or forty proprietors, with incomes answering to between one thousand and five thousand a year, would create a much more effective demand for wheaten bread, good meat, and manufactured products, than a single proprietor possessing a hundred thousand a year. (1989b, vol. 2, 373–4)

In his view, 'the division of landed property is one of the great means of the distribution of wealth', and without 'an easy subdivision of landed property ... a country with great natural resources might slumber for ages with an uncultivated soil, and a scanty yet starving population' (1989b, vol. 1, pp. 439–40).

He did not propose that either private property in land or the class of landed proprietors should be abolished. He regarded both as necessary. But he did not regard 'the present great inequality of property' as 'either necessary or useful to society'; and added that 'On the contrary, it must certainly be considered as an evil, and every institution that promotes it is essentially bad and impolitic' (Malthus 1986, vol. 1, p. 102). Less inequality in land ownership would mean that rents would be enjoyed by a larger number of proprietors.

However, he did not wish the division of property to be pushed too far:

> The division and distribution of property, which is so beneficial when carried only to a certain extent, is fatal to production when pushed to extremity. (1989a, I, 372)

This argument is an excellent illustration of his characteristic middle-way methodology. He himself regarded the question of the division of property as the most important application of the doctrine of proportions.

> It will be found, I believe, true that all the great results in political economy, respecting wealth, depend upon *proportions* ... But there is no part of the whole subject, where the efficacy of proportions in the production of wealth is so strikingly exemplified, as in the division of landed and other property; and where it is so very obvious that a division to a certain extent must be beneficial, and beyond a certain extent prejudicial to the increase of wealth. (1989b, vol. 1, pp. 432–3)

Distribution as a Factor of Production

Other writers, before and after Malthus, have discussed production and distribution, but generally their approach has been to regard distribution as the process whereby the proceeds of production are shared out after they have been produced by the factors of production. Their theory of distribution is worked out independently of their theory of production.

Malthus also looked at the problem of distribution in this way, with separate chapters analysing the way in which wages, profits and rents are determined. But in addition he looked at distribution from another direction. For Malthus, distribution is not merely concerned

with sharing out the spoils of production. It has a further function. It is an essential determinant of production, and an integral part of the production process considered in its totality. Without a proper distribution, there would be no production – except at a self-subsistence level. He saw distribution as a problem to be resolved *before* (as well as after) production takes place. Whereas others were concerned mainly with how the distribution of the product between wages, profits and rent is affected by economic development, Malthus made a major contribution by stressing that the distribution of the product in turn affects economic development. He was in effect saying, if not in these precise words, that distribution must be regarded as a factor of production, along with the conventional listing of the other factors of production – land, labour and capital. They represent only the supply side of the production process; but, if production is to occur, there must be a motive to produce as well as the means. In an exchange system, there will be no motive for producers to produce unless there are prospects of profits, and there will be no profit prospects unless there is an adequate effective demand for the products. This effective demand from potential consumers will not be forthcoming unless there has been a proper distribution of spending power. As Malthus said, 'there is certainly no indirect cause of production as powerful as consumption' (1989b, vol. 2, p. 34). The effective demand generated by a proper distribution provides demanders with the power or means to demand, and this provides suppliers with the will or motive to supply. It is obvious that, unless there has been production, there can be no distribution. But Malthus insisted that maximum production will not be achieved unless an optimum spread of distribution is established.

The separation and dichotomy between production and distribution that is presented in typical textbooks would therefore have been unacceptable to Malthus, for whom any listing of the factors of production would have to include distribution. It is this relationship of reciprocal causation between distribution and production that makes Malthus's theory of distribution innovative and distinctive.

Saving, Investment and Hoarding

Some commentators have interpreted Malthus as holding that savings are always invested; and have concluded that in Malthus's theory saving is not a leakage from the circular flow, does not constitute a reduction in effective demand, is not an impediment to economic growth, and must always be beneficial. In this respect, they see a major difference between Malthus and Keynes. They deny the claim that Malthus was a precursor of Keynes, and argue that Keynes was mistaken in regarding Malthus as a precursor. This interpretation appears to have been based in part on statements such as:

> it is stated by Adam Smith, and it must be allowed to be stated justly, that the produce which is annually saved is as regularly consumed as that which is annually spent, but that it is consumed by a different set of people. (Malthus 1989b, vol. 1, p. 31)

Malthus appears here to agree with Adam Smith that savings will always find an outlet in investment, and that there will never be a surplus of savings over investment. However, that interpretation is doubtful, given that 'and it must be allowed to be stated justly' was omitted from the second edition of the *Principles* (see Malthus 1989b, vol. 2, pp. 28, 300–1). Ricardo in his *Notes on Malthus* had said that a saving-equals-investment interpretation is inconsistent with the views expressed elsewhere by Malthus on saving. It would be reasonable to conclude that the omission was made by Malthus in response to Ricardo's note (Ricardo 1951–73, vol. 2, p. 15, n. 4).

Another possible source for attributing a saving-equals-investment view to Malthus might be Malthus's statement 'No political economist of the present day can by saving mean mere hoarding' (1989b, vol. 1, p. 32). Some commentators have interpreted this to mean that, in Malthus's view, savings are always invested, never hoarded, and never intended to be hoarded. If correct, such an interpretation would also constitute a major difference between Malthus and Keynes.

But there is another, more plausible, interpretation. Malthus here was not saying that savings are never held as idle cash balances. He was not precluding the possibility that savings might remain uninvested and idle, not on purpose but

because a satisfactory investment outlet cannot be found. This alternative interpretation negates a saving-equals-investment interpretation.

There are numerous instances in Malthus's writings that support this alternative interpretation. They clearly show that, in his view, savings will not always be invested, and that excessive savings are harmful. For example, Adam Smith had said that 'every frugal man [appears to be] a publick benefactor' and that the increase of wealth depends on a favourable balance of production over consumption (Smith 1776, book II, ch. 3, para. 25; book IV, ch. 3, iii, para. c.15), but Malthus disagreed:

> That these propositions are true to a great extent is perfectly unquestionable ... but it is quite obvious that they are not true to an indefinite extent. (1989b, vol. 1, 8)

To say that Malthus identified or equated saving and investment is to ignore his frequent use of expressions such as redundant capital, excessive capital, idle capital, spare capital, premature supply of capital, unemployed capital, vacant capital, capitalists at a loss where they can safely employ their capitals, capitals at a loss for employment, and so on. These expressions refer to funds that arise through savings and are intended for investment but for which an actual investment, at an acceptable degree of profit and risk, cannot be found. Malthus was thus recognizing the possible existence of an inequality between *ex ante* or intended investment and *ex post* or actual investment, because of the exhaustion of profitable investment outlets. This gap between savings intended for investment and savings actually invested could be described as unintended or residual hoarding – although Malthus did not use those terms – as distinct from the intended hoarding of a miser, or 'mere hoarding'.

Malthus and Ricardo

The correspondence between Malthus and Ricardo provides a revealing insight into the minds and characters of two of the most important contributors to the development of political economy in England during its formative years in the early 19th century (see Ricardo 1951–73, vols 6–9). They expressed their arguments forcefully but politely, although at times hints of frustration and exasperation began to appear, as they struggled to comprehend and to counter the other's point of view, especially when in an era without carbon copies and photocopies they seemed to forget what they had previously written. And despite their doctrinal and methodological differences, they remained close friends, with frequent visits to one another's homes. Ricardo's last letter to Malthus concluded with the statement:

'I should not like you more than I do if you agreed in opinion with me' (Ricardo 1951–73, vol. 9, p. 382); and Malthus, after the death of Ricardo, was reported to have said: 'I never loved any body out of my own family so much' (Empson 1837, p. 489). Ricardo had offered to assist Malthus financially by investing money for him in a stockbroking venture; and at one stage Malthus might have been seriously considering a personal involvement in international trading in commodities and bullion, using statistics and advice provided by Ricardo (see Malthus 2004, ch. 3). After Ricardo's death, Malthus defended him against critics who Malthus considered had gone too far in their criticisms; and, in lectures read to the Royal Society of Literature in 1825 and 1827, Malthus developed a theory of value which, while maintaining his previous emphasis on demand and supply as determinants of value, gave greater recognition to Ricardo's emphasis on the cost of production (Malthus 1986, vol. 7, pp. 301–23).

Opinions differ on who was the greater economist - Malthus or Ricardo. Who made the more significant contributions to the development of economics? On the one hand there are those who see Malthus as muddle-headed, and Ricardo as the better logician. On the other hand, there are those who reject the claim that Ricardo was a better logician, and who argue that Malthus's understanding of the multicausal complexity of the real world was of far greater value to the progress of economics than Ricardo's abstract theorizing. The most famous member of the latter group, J.M. Keynes, said that the world would be 'a much wiser and richer place' if 'Malthus, instead of Ricardo, had been the parent stem from which

nineteenth-century economics proceeded'; and that 'the almost total obliteration of Malthus's line of approach and the complete domination of Ricardo's for a period of a hundred years has been a disaster to the progress of economics' (Keynes 1933, pp. 120, 117).

See Also

- ▶ Corn Laws, Free Trade and Protectionism
- ▶ History of Economic Thought
- ▶ Keynes, John Maynard (1883–1946)
- ▶ Laissez-Faire, Economists and
- ▶ Malthusian Economy
- ▶ Smith, Adam (1723–1790)

Selected Works

1986. *The works of Thomas Robert Malthus*, 8 vols, ed. E. Wrigley and D. Souden. London: William Pickering.

1989a. *An essay on the principle of population*, variorum ed., 2 vols, ed. P. James. Cambridge: Cambridge University Press for the Royal Economic Society.

1989b. *Principles of political economy*, variorum ed., 2 vols, ed. J. Pullen. Cambridge: Cambridge University Press for the Royal Economic Society.

1997. *T.R. Malthus: The unpublished papers in the collection of Kanto Gakuen University*, vol. 1, ed. J. Pullen and T. Hughes Parry. Cambridge: Cambridge University Press for the Royal Economic Society.

2004. *T.R. Malthus: The unpublished papers in the collection of Kanto Gakuen University*, vol. 2, ed. J. Pullen and T. Hughes Parry. Cambridge: Cambridge University Press for the Royal Economic Society.

Bibliography

Bonar, J. 1924. *Malthus and his work*. 2nd ed. London: George Allen and Unwin.

Bonar, J. 1925–6. Thomas Robert Malthus. In *Palgrave's dictionary of political economy*. 2nd ed., ed. H. Higgs. London: Macmillan.

Dupâquier, J., A. Fauve-Chamoux, and E. Grebenik. 1903. *Malthus past and present*. London: Academic Press.

Empson, W. 1837. Life, writings and character of Mr Malthus. *Edinburgh Review* 64 (January): 496–506.

Ghosh, R. 1963. Malthus on emigration and colonization. Letters to Wilmot-Horton. *Economica* 30: 45–62.

Hollander, S. 1992. Malthus's abandonment of agricultural protectionism: A discovery in the history of economic thought. *American Economic Review* 82: 650–659.

Hollander, S. 1995. More on Malthus and agricultural protection. *History of Political Economy* 27: 531–537.

Hollander, S. 1997. *The economics of Thomas Robert Malthus*. Toronto: University of Toronto Press.

James, P. 1966. *The travel diaries of Thomas Robert Malthus*. Cambridge: Cambridge University Press.

James, P. 1979. *Population Malthus: His life and times*. London: Routledge and Kegan Paul.

Keynes, J.M. 1933. Robert Malthus . In *Essays in biography*. London: Macmillan, 1961.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

McCleary, G. 1953. *The Malthusian population theory*. London: Faber and Faber.

Meek, R., ed. 1953. *Marx and Engels on Malthus*. London: Lawrence and Wishart.

Otter, W. 1836. Memoir of Robert Malthus. In *Principles of political economy*. 2nd ed., ed. T.R. Malthus. New York: Augustus M. Kelley, 1951.

Pullen, J. 1987. Malthus, Thomas Robert. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London: Macmillan.

Pullen, J. 1995. Malthus on agricultural protection: An alternative view. *History of Political Economy* 27: 517–529.

Pullen, J. 2004. Malthus, Thomas Robert. In *Oxford dictionary of national biography*. Oxford: Oxford University Press.

Ricardo, D. 1951–73. *The works and correspondence of David Ricardo*, 11 vols, ed. P. Sraffa with the collaboration of M. Dobb. Cambridge: Cambridge University Press for the Royal Economic Society.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. R. Campbell and A. Skinner; textual editor W. Todd. Oxford: Clarendon Press, 1976.

Stephen, L. 1893. Malthus, Thomas Robert. In *Dictionary of national biography*, vol. 12. London: Oxford University Press.

Waterman, A. 1991. *Revolution, economics and religion: Christian political economy, 1798–1833*. Cambridge: Cambridge University Press.

Winch, D. 1983. Higher maxims: Happiness versus wealth in Malthus and Ricardo. In *That noble science of politics*, ed. S. Collini, D. Winch, and J. Burrow. Cambridge: Cambridge University Press.

Winch, D. 1987. *Malthus*. Oxford: Oxford University Press.

Winch, D. 1996. *Riches and poverty: An intellectual history of political economy in Britain, 1750–1834*, Part 3, Robert Malthus as political moralist. Cambridge: Cambridge University Press.

Malthus's Theory of Population

D. R. Weir

It is often said that Malthus was a better historian than prophet. He would be disappointed in that verdict because his intention in formulating his Theory of Population was to create a scientific basis for predicting the future state of mankind, in opposition to the speculations of utopian writers, especially Godwin. His failure to predict the Industrial Revolution is the evidence most often brought against him. But even with the benefit of hindsight, economic historians today still find it difficult to predict the Industrial Revolution from its antecedents. The crucial contribution of Malthus's theory was not its pessimism about innovation but rather its prediction of the demographic consequences of technological change and the inevitable effect of population on the standard of living. Malthus's Theory of Population continues to influence economic thought from popular discussion to policy-making, to model-building – long after many of its classical contemporaries, like the labour theory of value, have passed from the scene.

This essay will focus on the population side of Malthus's theory. We begin with a distillation of his ideas into a simple model in the modern sense of the term. Our intention is not to treat in detail all the nuanced and sometimes contradictory aspects of Malthus's writing. The *Essay on Population*, first published in 1798, was revised six times before the final seventh edition was published in 1872, some 38 years after his death. The model used here aims to portray the most essential and durable aspects of the theory. It also provides an organizational framework for discussing the evidence for and against its predictions from time periods both before and after Malthus wrote.

The Model

Figure 1 portrays the essential elements of Malthusian equilibrium. There are three curves, representing three functional relationships. In the first panel is an aggregate production function showing the standard of living (or real wage, or income per capita) produced by a population of a given size. Its main feature is diminishing returns to labour – a tenet of classical economics not unique to Malthus. The second panel describes demographic behaviour. Mortality (here, the crude death rate, which is the number of deaths per 1,000 persons) rises as the standard of living falls. This is the positive check. Fertility (here, the crude birth rate), falls as the standard of living falls. This is the preventive check. Population grows when births exceed deaths and falls when deaths exceed births. A rising population lowers the standard of living (through the production function), which in turn raises mortality and lowers fertility, eventually bringing population growth to a halt. Equilibrium in this simple version of the model is attained at zero population growth. At that point, wages do not change, and consequently the birth and death rates do not change. The equilibrium is stable, since any disturbance sets in motion compensating changes.

The stability of the equilibrium is the source of Malthusian pessimism. Imagine an expansion of land area for cultivation. The production function would shift out, raising the standard of living for the current population. Fertility would rise and mortality fall; population growth would continue to devour the gains until the wage fell to its original level. Demographic behaviour is the forge of the Iron Law of Wages. Permanent change in the standard of living can arise only from restraint of fertility (a lower birth rate at each wage) or a worsening of mortality (more deaths at each wage).

The smooth curves drawn above describe the long-run tendencies as envisaged by Malthus. He saw the process of adjustment, however, as anything but smooth. Population growth would tend to overshoot the equilibrium. The positive check, working through disasters like major famines or disease, would be slow to respond but would then overadjust when it did, setting off a new cycle. Malthus offered no specifics on the periodicity or amplitude of the cycle, only the prediction of oscillations around a long-run equilibrium level.

The Evidence

In discussing in detail the component parts of the model, there are three aspects of each to be considered. First if the evolution of Malthus's own ideas on the functional relationship; second, the historical evidence for or against it; and third, its relevance to the major economic and demographic transformations of the last two centuries. The importance of empirical verification was strongly emphasized by Malthus himself. Criticizing the utopians, Malthus wrote:

> A writer may tell me that he thinks man will ultimately become an ostrich. I cannot properly contradict him. But before he can expect to bring any reasonable person over to his opinion, he ought to show that the necks of mankind have been gradually elongating, that the lips have grown harder and more prominent, that the legs and feet are daily altering their shape, and that the hair is beginning to change into stubs of feathers (Malthus 1798).

Much of the work of the later editions of the '*Essay on Population*' was devoted to amassing evidence on the Principle of Population at work.

A central problem with using this or any other equilibrium framework to explain co-variations in economic and demographic variables (across time or space), or with using empirical observations to 'test' the model, is the identification of exogenous shocks as distinct from endogenous responses. Malthus himself was vaguely aware of the problem as early as the first edition of his *Essay*. David Hume, noting the early age at marriage of women in China, had deduced that the population must as a consequence be very large. Malthus, taking age at marriage as endogenous rather than exogenous, concluded that the population must on the contrary be rather small and wages relatively high to induce such early marriage.

Mortality

Malthus gave to mortality's response to wages the name 'positive' check because it was certain and unavoidable once population had grown too large. Fertility offered a 'preventive' check in the sense that if low fertility held back the growth of population, the mortality response could be postponed.

Malthus envisaged two modes of action for the positive check. Associated with declining wages would be increasing 'misery and vice'. Misery and vice included some conditions that would raise mortality as well as lower mortality. This would yield a smooth continuous relationship like that in the diagram. The second mode of action would be sudden mortality crises to greatly reduce a population in a year or two. To put it in modern terms, the probability of a mortality crisis of any given magnitude should increase as the standard of living falls. The expected value of the death rate would show the smooth relationship to wages pictured above. Actual events would be much less regular. Adjustment to equilibrium was inevitable but not constant.

Historical studies of the positive check have approached it from two very different perspectives. One sort examines great crises to determine whether they resulted from population pressure. The other attempts to specify the extent of population pressure on resources over time and look for mortality consequences.

The Black Death of 1347 and 1348 killed between one-third and one-half of Europe's population in a single massive epidemic of bubonic plague. Hatcher (1977) concludes for England that the Black Death was 'not Malthusian', by which he means not a response to population pressure. Evidence abounds that population had been growing: rents and the relative price of basic foods had been rising for two centuries or more. There had been major famines in the 1320s. The 'anti-Malthusian' conclusion is based on the absence of a logical connection between standard of living and the scale of epidemic bubonic plague and on the fact that the mortality response was disproportionate to the population pressure. Economic responses to the Black Death were clearly Malthusian (Hatcher 1977, pp. 101–94). It is certainly consistent with the subtler Malthus to find that an induced (endogenous) mortality response would overreact and become an exogenous disturbance driving population below its equilibrium. The study of a single episode cannot determine whether the probability of its occurrence was raised by the economic conditions preceding it. The persistence of the plague and

continued population declines of the next century or more, clearly were not a consequence of the improved standard of living after the first outbreak.

Half a millennium later the Irish Potato Famine provided a tragic forum for debate over the new Malthusian ideas. Mokyr (1983) traces reliance on the potato to population pressure, through the unique funnel of Irish institutions. As was the case with the plague, the element of chance looms large in the timing and location of the potato blight itself. Unlike the Black Death, however, the means were available to alleviate the heavy mortality consequences. One wonders what might have been the fate of the Irish had not English policy-makers of the 1840s been educated in the science of Malthus's Theory of Population.

The main supporting evidence for the positive check comes from the study of subsistence crises: short-run mortality increases following harvest failures. Meuvret (1946) drew attention to the close association of grain prices and deaths in specific incidents in France. Subsequent studies have shown a regular statistical association over long periods in several countries. Improved marketing and production methods appear to have been successful in nearly eliminating the relationship in England by the end of the 16th century and in France by the middle of the 18th century.

Since Malthus wrote, life expectancy has increased from well under 40 to well over 70 in most of the now developed world. Since the standard of living has also increased, it would appear that Malthus's theory of mortality has been a better prediction than it was a description of prior history. Some scepticism on that point is voiced by Preston (1976), who finds that life expectancy across countries is not closely related to their level of per capita income at any point in time. He concludes that medical and public health technology are more important than income. Since medical technology may well be a function of per capita income in the leading country, or in the average of leading countries, his finding may indicate only that Malthus no longer applies within national boundaries.

There is, in sum, only limited evidence of the income–mortality relation postulated by Malthus. Evidence abounds that most of the variation in mortality cannot be accounted for by so simple a framework.

Fertility

In the first edition of the *Essay on Population* Malthus claimed that 'passion between the sexes was necessary and will remain nearly in its present state'; that is, that fertility was roughly constant and did not vary with living standards. On our diagram, the fertility curve would be vertical. He subsequently advocated delayed marriage as a check to population growth. Being a vicar of the Anglican Church, he condemned both contraception and never-marrying as 'vice'. It is one of the greatest misattributions in human history that associates the name of Malthus with the birth-control movements of the late 19th century. Nevertheless, since fertility restraint offers the only means of raising both the standard of living and length of life within his system, it is not surprising that those who believed his model but not his morals would eventually invoke his science to promote their cause.

The evidence for endogenous fertility responses is growing but is not yet convincing. Wrigley and Schofield (1981) find that long-run trends in English fertility followed long-run trends in real wages from 1541 to 1871. The lag, however, was 40 years, and the two moved in opposite directions for approximately 140 of the 330 years. Moreover, in England, as in the rest of early modern Europe, age at marriage and fertility within marriage were fairly constant, leaving changes in proportions ever-married as the main source of changes in English fertility before the Industrial Revolution (Weir 1984).

The greatest failure of Malthus's Theory of Population is in explaining the fertility transition from high, mostly uncontrolled fertility within marriage to modern low fertility. The process began first in France at the time Malthus was writing. Parts of the United States and Hungary also began at about that time. The rest of Europe

followed sometime between 1870 and 1914. In no case was the long-run downward trend in fertility caused by a downward trend in national income. In today's developing countries fertility decline has sometimes been induced by policy measures without economic development, but sustained economic growth continues to be a prescription for contraception.

Neoclassical theories of fertility, as in the work of Becker (1981), salvage Malthus's theory as an income effect in a model of the demand for children. Substitution effects from a rising price of children relative to other consumption goods may be a more important determinant of fertility trends during development, overwhelming the income effect. Such a model does help explain why the Malthusian model seems to explain cyclical fluctuations in fertility both before (Lee 1981) and after (Easterlin 1973) the fertility transition. Relative prices may not fluctuate as much as income. Marx, a harsh critic of Malthus, would not be surprised. He claimed that Malthus's laws of population were specific to the particular mode of production of pre-industrial Europe. Other modes of production would have other modes of reproduction. Unfortunately, he left no better guide to predicting the changes than did Malthus.

It is perhaps ironic that Malthus's Theory of Population, conceived as a prediction of long-run equilibrium, should be consistent with short-run fluctuations but not with long-run movements. Humankind has not reached the idyllic state anticipated by Malthus's utopian adversaries. Neither have we fulfilled Malthusian predictions. The separation of reproduction from 'passion between the sexes' has led the wealthiest of nations to fertility below the level needed to replace their populations while in the poorest of nations great numbers of children are born into short lives of poverty. Theories of population must acknowledge their debt to Malthus and move on.

See Also

- ▶ Corn Model
- ▶ Demographic Transition
- ▶ Exhaustible Resources
- ▶ Limits to Growth
- ▶ Natural Resources

Bibliography

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

Easterlin, R.A. 1973. Relative economic status and the American fertility swing. In *Family economic behavior: Problems and prospects*, ed. E.B. Sheldon. Philadelphia: J.B. Lippincott.

Hatcher, J. 1977. *Plague, population, and the english economy, 1348–1530*. London: Cambridge University Press.

Lee, R.D. 1981. Short-term variation: Vital rates, prices, and weather. In Wrigley and Schofield (1981), ch. 9.

Malthus, T.R. 1798. *An essay on the principle of population, as it affects the future improvement of society. With remarks on the speculations of Mr. Godwin, M. Condorcet, and other writers*. Harmondsworth: Penguin, 1970.

Meuvret, J. 1946. Les crises des subsistances et la démographie de la France de l'Ancien Régime. *Population* 1(4): 643–650.

Mokyr, J. 1983. *Why Ireland starved*. London: Allen & Unwin.

Preston, S.H. 1976. *Mortality patterns in national populations*. New York: Academic Press.

Weir, D.R. 1984. Rather never than late: Celibacy and age at marriage in English cohort fertility, 1541–1871. *Journal of Family History* 9(4): 340–354.

Wrigley, E.A., and R. Schofield. 1981. *The population history of England, 1541–1871: A reconstruction*. Cambridge, MA: Harvard University Press.

Malthusian Economy

Gregory Clark

Abstract

The Malthusian economy was the economic system that characterized almost all economies before the industrial revolution. In this regime fertility and mortality rates at different material income levels determined the average real income level and life expectancy at birth. Thus before 1800 the improvement of production technologies resulted only in population growth, and not in any gains in material living

conditions beyond those that were found in the original hunter gatherer societies.

Keywords

Anthropometric history; Birth and death rates; Fertility; Historical demography; Industrial revolution; Malthus, T.; Malthusian economy; Population growth; Technology schedule

JEL Classifications

N3

The Malthusian economy is the economic system which prevails whenever a society's production technology advances so slowly that population growth forces incomes down to the subsistence level. In such an economy material welfare is independent of natural resources, technology and capital accumulation, but instead depends solely on the factors governing fertility and mortality. The resulting subsistence income can, however, vary widely across societies. Some Malthusian economies were rich by the standards of most countries in modern Africa, for example.

Almost all societies until 1800 were Malthusian, from the original foragers of the African savannah 50,000 years ago down through settled agrarian societies of considerable sophistication such as England, France, China and Japan in 1800. The operation of all human societies through history up until the Industrial Revolution can thus seemingly be described by this one simple economic system. An implication of this is that there was most likely no gain in material welfare between the evolution of anatomically modern humans and the onset of the Industrial Revolution.

Government actions, in so far as they change fertility or mortality, can influence material welfare in the Malthusian economy, but in a contradictory fashion. Good governments that reduced mortality through order and security made people poorer. Bad governments that increased mortality through warfare and banditry made them wealthier.

The economic logic of these societies was first, though only partially, appreciated by Thomas Malthus in his famous *Essay on a Principle of Population* of Malthus 1798. Malthus's insights were elaborated by writers such as David Ricardo and James Stuart Mill into the system called classical political economy in the early 19th century. Ironically, this intellectual development happened just as for the first time the rate of technological advance was becoming sufficiently rapid to bring the Malthusian era to a close.

Insight into the Malthusian economy starts from the insight that the biological capacity of women to produce offspring is much greater than the number of births required to reproduce the population. If fertility is unrestricted women can have 12 or more children. Social institutions regulating marriage and contraceptive practices will determine the actual numbers of births per women. In modern societies these institutions and practices vary greatly, so the number of births per women varies greatly. Completed fertility now ranges across the world from a low of 1.15 in Spain to a high of 8.0 in Niger. Only where women happen on average to have two children who survive to adulthood will population be stable. Even small deviations from this number will cause rapid increases or decreases in population. Thus modern populations are not stable.

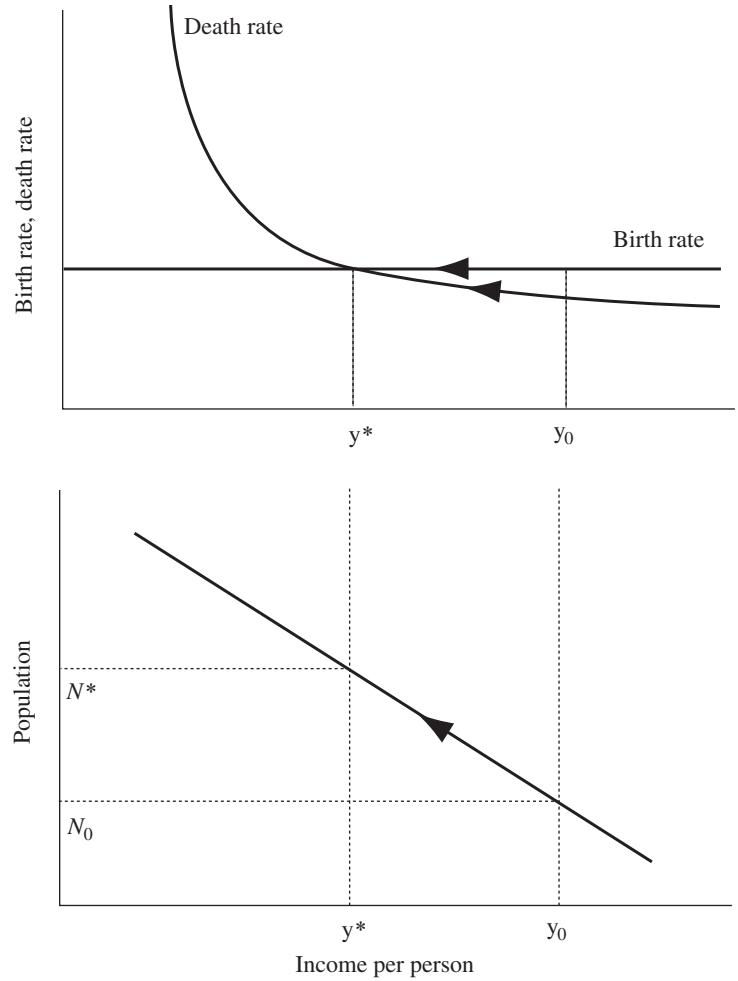
Despite this potential for explosive population growth, pre-industrial populations were remarkably stable over the long run. The average annual growth rate of world population from 10,000 years BC to AD 1800 was 0.05 per cent. The typical woman before 1800 thus had 2.02 children who survived to reproductive age. As an extreme case the population of Egypt, for example, is estimated at between four million and five million at 1000 years BC. The population in Greek and Roman Egypt a millennium later is estimated at this same four million to five million. The first modern census in 1848 suggests a population of 4.5 million. Thus over nearly 3,000 years the Egyptian population growth rate was to a close approximation zero, and women on average had two surviving children. Yet it is estimated that in Roman Egypt the average woman gave birth to six children. Some mechanism kept fertility and mortality in balance in these pre-industrial economies.

The Malthusian Equilibrium

The simple Malthusian model of how pre-industrial society functioned supplies an economic mechanism to explain its population stability. In its simplest version there are just three assumptions:

1. The *birth rate*, the number of births per year per thousand people, is a socially determined constant, independent of material living standards. Birth rates will vary across societies, but in this simplest model they are assumed to be independent in any given society of material living conditions.
2. The *death rate*, the number of deaths per year per thousand persons, declines as material living standards increase. Again, the death rate will differ across societies depending on climate and lifestyles, but it assumed that in all societies it will decline as material living conditions improve.
3. Material *living standards* decline as population increases.

Figure 1 shows the first two assumptions of the simple Malthusian model in graphical form in the upper panel. The birth and death rates are plotted on the vertical axis, material income per capita on the horizontal axis. The first two assumptions of the simple Malthusian model imply that there is only one level of real incomes at which the birth rate equals the death rate, denoted as y^*. And this constitutes a stable equilibrium. Thus y^* is called



Malthusian Economy, Fig. 1 Long-run equilibrium in the Malthusian economy

the 'subsistence income' of the society: it is the income at which the population barely subsists, in the sense of just reproducing itself. This subsistence income is determined without any reference to the production technology. It depends only on the factors which determine birth and death rates. Once we know these factors we can determine the subsistence income.

Another aspect of human welfare is life expectancy at birth, that is, the average number of years a person will live. In the Malthusian era life expectancy at birth also depended only on the factors determining birth and death rates. This is because with a *stable* population, where annual births have equalled deaths for a long time, life expectancy at birth is the inverse of the crude birth rate. With fertility not restricted in any way crude birth rates would be 50–60 per thousand in pre-industrial populations (based on modern experience). This would imply a life expectancy at birth of 20 years or less.

The term 'subsistence income' can lead to the confused notion that in the Malthusian economy people were always living on the edge of starvation. In fact, in almost all Malthusian economies the subsistence income was considerably above the income required for the physiological minimum daily diet. All pre-industrial societies for which we have good demographic records limited fertility below the biological maximum. Differences in the location of the mortality and fertility schedules generated subsistence incomes at very different levels. Thus, both 1450 and 1650 were periods of population stability in England, and hence periods where by definition income was at subsistence. But the wage of unskilled agricultural labourers was equivalent to about 6 lb of wheat flour per day in 1650, compared with 18 lb in 1450. Even the 1650 unskilled wage was well above the physiological minimum. A diet of about 1.33 lb of wheat flour per day would keep a labourer alive and fit for work (it would supply about 2,400 calories per day). Thus, pre-industrial societies, while they were subsistence societies, were not starvation regimes. England in 1450, indeed, was wealthy even by the standards of many modern societies such as those in sub-Saharan Africa.

The bottom panel of Fig. 1 illustrates the third assumption. The panel has on the vertical axis the population, N, and on the horizontal axis the material income. As population increased material income per person by assumption declined. The justification for this assumption is the law of diminishing returns. Since one important factor of production, land, is always in fixed supply in pre-industrial economies, the law of diminishing returns implies that average output per worker fell as the labour supply increased as long as the technology remained static. Thus the average amount of material consumption available per person fell with population.

Figure 1 also shows how an equilibrium birth rate, death rate, population level and real income were arrived at in the long run in a pre-industrial economy. Suppose we start at an arbitrary initial population N_0 in the diagram, greater than N^*. This generates an income y_0, above the subsistence income. At this income the birth rate exceeds the death rate, so population grows until income falls to y^* and population equals N^*.

Changes in the Birth Rate, Death Rate and 'Technology' Schedules

Suppose that the birth rate schedule in Fig. 1 was higher. Then at the equilibrium, real income would be lower, and the population greater. Thus any increase in birth rates in the Malthusian world drove down real incomes and reduced life expectancy. Conversely, anything which limited birth rates drove up real incomes and increased life expectancy. Thus in the pre-industrial era birth rates were a crucial determinant of material living conditions.

If the death rate schedule was higher, so that at each income there was a higher death rate, then the equilibrium real income would be higher. But if the birth rate was not responsive to income then a greater death rate increased real incomes but in the long run had no effect on the annual death rate or on life expectancy at birth.

Thus in this simplest Malthusian model higher mortality risks at a given income were unambiguously a good thing, at least in the long run.

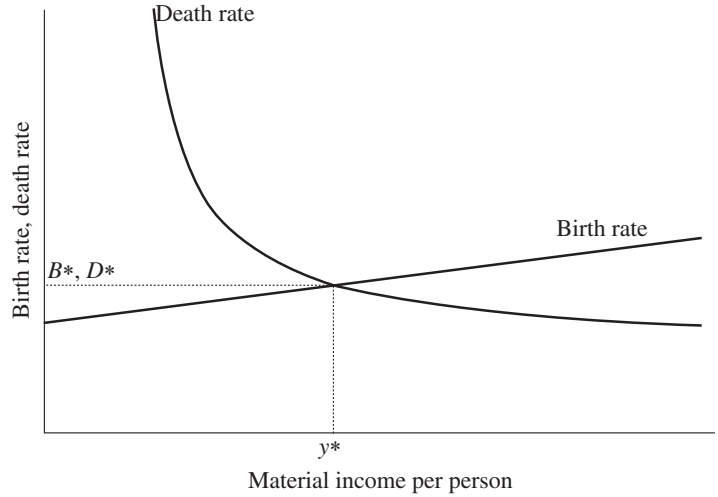
The simple Malthusian world thus exhibits an almost counter-intuitive logic. Anything that raised the death rate schedule, the death rate at a given income, such as war, disorder, disease or poor sanitary practices, increased material living standards without changing life expectancy at birth. Anything that reduced the death rate schedule, such as advances in medical technology, or better public sanitation, or public provision for harvest failures, or peace, reduced material living standards without any gain in life expectancy at birth.

While the real income was determined from the birth and death schedules, the population size depended on the schedule linking population and real incomes. Above I labelled this the 'technology' schedule, because in general the major cause of changes in this schedule has been technological advances. But other things could shift this schedule – a larger capital stock, improvements in the terms of trade, climate improvements, and a more productive organization of the economy. A shift upwards in this schedule, in the short run, since population can change only slowly, would have increased real incomes. But the increased real incomes reduced the death rate, so that births exceeded deaths and population began growing. The growth of population ended only when the income returned to the subsistence level, y^*. At this new equilibrium the only effect of the technological change was to increase the population supported. There was no lasting change in the living standards of the average person.

More Complicated Malthusian Models

An issue that has exercised historical demographers is whether the birth rate in pre-industrial societies was 'self-regulating'. What they mean by this is shown in Fig. 2, which shows the birth and death schedules of a simplified Malthusian model, as well as a modified birth schedule, which slopes upwards with material incomes. In the modified Malthusian model it is assumed that in good times people married earlier and more people married, so that fertility increased, whereas in bad times fewer married, and they married later, so that fertility declined.

It should be clear that a positive association of fertility and income does not change the basic equilibrium of the model. The only difference is that increases in the death rate at any given material income are now not so unambiguously good, since they will be associated with higher fertility and mortality rates and hence lower incomes. The evidence for societies such as pre-industrial England, however, shows no response of fertility to income (Wrigley et al. 1997). Thus the simple model may well describe pre-industrial societies well.



Malthusian Economy, Fig. 2 A Malthusian model where births increase with income

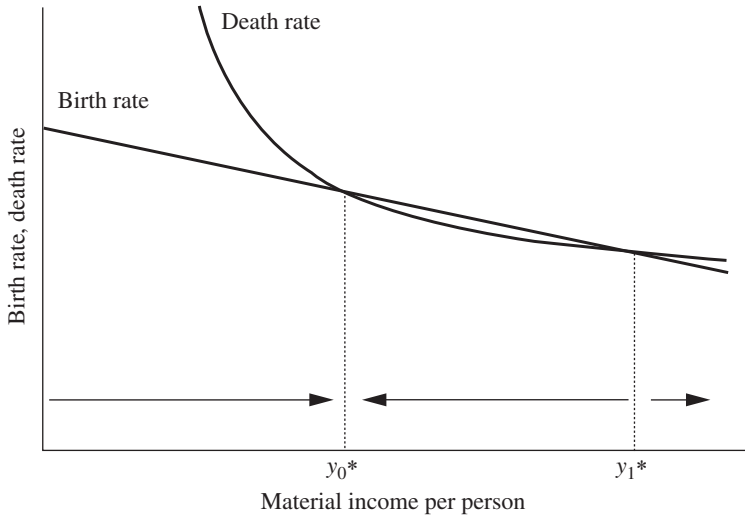
What causes many more potential complications is a birth schedule that declines with material incomes. Suppose that as real incomes go up one of the responses of people is to desire fewer children. With a birth rate that declines with real incomes the model could have multiple crossings between the birth rate and death rate schedules. At those places where the birth rate schedule was declining more steeply than the death rate schedule the equilibrium would be unstable. Figure 3 gives a declining birth rate schedule that twice intersects the death rate schedule. The intersection at the lower real income, y_0, is a stable equilibrium. But the second higher income equilibrium at y_1 is unstable. If real incomes drop below this level by any amount then population starts to grow, leading real incomes all the way down to the stable equilibrium at y_0. Conversely if they increase at all above y_1 then deaths will exceed births and real incomes continue to grow indefinitely. The population will fall eventually to zero.

In this case there is a 'Malthusian trap' in the pre-industrial economy. A society can be stuck in the subsistence income equilibrium unless some jolt such as acquiring extra land, experiencing a much higher death rate, or experiencing faster technological progress pushes up wages enough so that fertility falls permanently. The shock of the Black Death, however, which tripled real incomes for the poorest workers in England by 1450, did not lead to any permanent movement towards lower fertility and the escape from the Malthusian trap. Again, the evidence for pre-industrial demography suggests no declines in fertility with higher incomes.

The Empirical Implications of the Malthusian Model

The most interesting empirical implication of the Malthusian model is that material living conditions for people, including life expectancy at birth, may well have been unchanged between the dawn of humanity and AD 1800. Were the people in sophisticated societies such as England, France, the Netherlands, Japan and China in 1800 really no better off than the original hunter-gatherers? This seems particularly counter-intuitive for England, reckoned to be the richest country the world by 1800.

By then England was a society that would not seem that different from our own. The middle and upper classes in London breakfasted at coffee shops as they read the daily newspapers. They dwelled in homes of brick and glass with water supplied by lead pipes, lighted at night by oil harvested from sperm whales taken thousands of miles away in the oceans. There was extensive trade for luxury products from the tropics – cottons, silks, spices. How could the material condition of humanity not be better then than in the



Malthusian Economy, Fig. 3 A Malthusian model where births decline with income

savage past when our ancestors faced the elements naked, and sought shelter at night in depressions in the ground or in crude lean-tos?

But even in England in 1800 the living conditions of the mass of the population were still primitive. The largest employment was still agriculture, where the average day wage in 1800–9 was the equivalent of 5.7 lb of wheat flour. This was enough to keep a family fed only if most of the income was spent on the cheapest forms of food such as bread. Farm labourers lived in simple structures little better than those of the medieval period. They slept when it was dark because they could not afford lighting. They could afford one new set of clothing per year. English farm labourers six hundred years before, in 1200–9, received a wage which was the equivalent of 12 lb of flour, significantly more than in 1800. And at the best time for pre-industrial workers in England, circa 1450, when the population losses of the plagues which ravaged Europe from 1348 on were their greatest, the real wage was much higher, equivalent to 18 lb of flour. In the years 1200–1800 in England there is no sign of long-run gains in real wages for the mass of workers. We know also the real day wage of farm workers in Roman Egypt circa AD 250 was the equivalent of 5 lb of flour, not much less than England in 1800.

How did English material living conditions around 1800 compare with hunter-gatherer societies such as those that constituted society through the great bulk of human history? We can obtain insight on this in two ways. The first is by comparing living conditions in England in 1800 with those of the few surviving hunter-gatherer groups. Since the diets were very different here we have to use measures such as the number of calories consumed per person per day. In 1787–96 for the families of English farm workers this was a meagre 1,508 calories. For a group of eight hunter-gatherer societies studied in the 1960s to 1980s the average consumption was 2,272 calories, much better than for England. On this measure the English on the eve of the Industrial Revolution seem to have lived less well than the average hunter-gatherer. Another aspect of the quality of life is life expectancy at birth. One measure of this is the fraction of infants that survived the first year of life. In England as a whole this is estimated at 83 per cent in the second half of the 18th century. For modern hunter-gatherer societies survival rates were a little lower at 79 per cent. But this is still not that much lower than for the richest society in the world in 1800. And survival rates for infants in London, the richest part of England, were only 70 per cent because of the health hazards of city life.

A second measure is the average stature of people. Height is a good index of material living conditions, since it depends on both food consumption and the amount of sickness people experience as they grow. Average heights for adult males in England circa 1800 were 67 inches or less. This was very good by the standards of societies just before industrialization. Average male heights in Japan in the late 19th century were 61 inches and in India in the early 19th century 64 inches. Yet these heights in England are little if any better than those recorded from skeletons of hunter-gatherers in the Mesolithic (10,000–5000 BC) and Neolithic (5000–1000 BC) in Europe. Average male height from these skeletons is estimated at 66 inches. So overall, if we look at agrarian societies across the world in 1800 AD, the stature evidence suggests a decline in living conditions from hunter-gatherer society.

Thus, the evidence is that for the mass of humanity on the eve of the Industrial Revolution living conditions were no better and probably worse than in the hunter-gatherer past.

See Also

- ▶ Anthropometric History
- ▶ Historical Demography
- ▶ Industrial Revolution
- ▶ Malthus, Thomas Robert (1766–1834)

Bibliography

Bennike, P. 1985. *Paleopathology of danish skeletons*. Copenhagen: Akademisk Forlag.

Clark, G. 2007. *A farewell to alms: A brief economic history of the world*. Princeton: Princeton University Press.

Koepke, N., and J. Baten. 2005. The biological standard of living in Europe during the last two millennia. *European Review of Economic History* 9: 61–95.
Malthus, T. 1798. *An essay on a principle of population*. 6th ed, 1826. London: John Murray.
Miller, M., and C. Upton. 1986. *Macroeconomics: A neoclassical introduction*. Chicago: University of Chicago Press.
Steckel, R., and J. Rose, ed. 2001. *The backbone of history: Health and nutrition in the Western Hemisphere*. Cambridge: Cambridge University Press.
Wrigley, E., R. Davies, J. Oeppen, and R. Schofield. 1997. *English population history from family reconstruction: 1580–1837*. Cambridge, MA/New York: Cambridge University Press.

Malynes, Gerard de (*fl.* 1586–1623)

Lawrence H. Officer

Keywords
Exchange controls; Malynes, G. de; Money supply; Price level; Specie-flow mechanism; Terms of trade; Usury

JEL Classifications
B31

A merchant of English parentage, born in Antwerp at an unknown date, Malynes was a commissioner of trade in the Low Countries about 1586. He came to London and was frequently consulted on commercial questions by the Privy Council in the reigns of Elizabeth I and James I. He became an assay master at the mint and obtained a patent to supply farthings; he was imprisoned for a time, complaining later that he had been ruined by being paid in his own coins. He also served as a spy for England. Called on by the standing commission on trade for evidence on the state of the coinage, he published a series of pamphlets on money and prices. A mercantilist and a bullionist, he was heavily influenced by Scholastic literature.

Malynes viewed individual commodity prices as determined by demand and supply. However, he was more interested in the price level, governed by the quantity of money (Malynes 1601b, 1603). An expanding money supply, associated with a rising price level, decreased interest rates and stimulated the economy (1601b, 1622a). Therefore Malynes viewed usury as at best a necessary evil (see Muchmore 1969, p. 346) and, above all, opposed any export of specie whatsoever.

Rejecting the balance of trade theory, Malynes charged that 'bankers' (exchange dealers) controlled the exchange rate (1601b, 1622a, b, 1623). By their incorporation of usury in the price of a bill of exchange and through speculation, they conspired to undervalue sterling, leading to a deterioration in England's terms of trade ('overbalancing') and a specie outflow (1601b, 1622a, 1623). But overvalued sterling would not lead to a specie inflow, because the export proceeds would be spent on luxury imports (1601b). Yet Malynes (1601b) has a theory of price level changes in response to exchange rates differing from mint parity and money flowing between countries – a price specie-flow mechanism, marred only by the assumption of inelastic demand. His solution to the twin problems of specie outflow and terms of trade deterioration is comprehensive exchange control with enforced exchange dealings at rates fixed at mint parities (Malynes 1601b, 1622a, b; Muchmore 1969, pp. 347–8).

Selected Works

1601a. *Saint George for England, allegorically described*. London: Richard Field for William Tymme.
1601b. *A treatise of the canker of England's commonwealth*. London: Richard Field for William Iohnes. Reprinted in part in *Tudor economic documents*, vol. 3, ed. R.H. Tawney and E. Power. London: Longmans, Green, 1924.
1603. *England's view, in the unmasking of two paradoxes*. London: Richard Field.
1622a. *Consuetudo, vel lex mercatoria, or the Ancient Law-merchant*. London: Adam Islip.
1622b. *The maintenance of free trade*. London: I. Legatt for W. Sheffard.
1623. *The centre of the circle of commerce*. London: William Iones.

Bibliography

Johnson, E. 1933. Gerard de Malynes and the theory of the foreign exchanges. *American Economic Review* 22: 441–455.
Muchmore, L. 1969. Gerrard de Malynes and mercantile economics. *History of Political Economy* 1: 336–358.
Officer, L.H. 1982. The purchasing-power-parity theory of Gerrard de Malynes. *History of Political Economy* 14: 251–255.
de Roover, R. 1974. *Business, banking and economic thought*. Chicago: University of Chicago Press.
Schumpeter, J. 1954. *History of economic analysis*. Oxford: Oxford University Press.
Spiegel, H. 1971. *The growth of economic thought*. Englewood Cliffs: Prentice-Hall.
Wu, C.-Y. 1939. *An outline of international price theories*. London: George Routledge & Sons.

Managerial Capitalism

Alan Hughes

As private enterprise industrial economies evolve, the proponents of the thesis of managerial capitalism argue that changes occur in the technical conditions, and scale, of production of corporations; in the structure of the ownership of their equity, and of the product markets in which they operate; and in their internal governance. The increasingly complex technological and scientific nature of industrial production requires specialist technical management expertise, and management responsibility is delegated to individuals who possess it by an increasingly absentee ownership interest. The associated increase in the scale and capital intensity of production is reflected both in the growth of oligopolistic market structures and, as external funding increases, by an ever more dispersed pattern of share ownership. Thus, the fusion between management and ownership is broken. Those responsible for exercising management responsibility come to constitute a skilled, inside, professional salaried group, essentially propertyless in relation to their own corporations, and hence separate in function and identity from the tens of thousands of shareholders who are its legal proprietors. These share owners, in turn, form an increasingly disparate, unorganized, and uninterested group of principals, unwilling or unable to impose their own self-interested contractual conditions of employment on their manager-agents. Managerial behaviour is therefore discretionary behaviour very weakly constrained by shareholder-owner interests on the one hand, or by a competitive market environment on the other. As a result, corporate behaviour changes, and with it so does the nature of capitalism (Veblen 1924; Berle and Means 1932; Berle 1955, 1960; Dahrendorf 1959; Marris 1964; Williamson 1964; Nichols 1969).

Whilst writers in the managerialist tradition agree that the separation of ownership from control has occurred and does matter, there is disagreement on the nature of the changes it produces. Berle argued that the most likely outcome would be a more socially responsive and socially responsible form of corporate behaviour, a vision shared by a number of other post-war authors (Bell 1961, 1974; Berle 1955, 1960; Drucker 1951; Mayo 1949; Mason 1960). On this view the development of a managerial corporate conscience would ensure behaviour responsive to public opinion and to a wider constituency of interests than the owners of capital alone. This would ensure the acceptability and survival of the corporate system. In the words of Berle and Means

> It is conceivable – indeed it seems almost essential if the corporate system is to survive – that the 'control' of the great corporations should develop into a purely neutral technocracy, balancing a variety of claims by various groups in the community and assigning to each a portion of the income stream on the basis of public policy rather than private cupidity (1932, p. 356).

This pluralist, or in the terminology of Nichols (1969), 'non-sectional' interpretation of the impact of managerialism has formed part of a wider stream of thought mapping out the socio-political and economic development of industrial societies. This has included as one possible outcome, resulting from similar underlying managerial and technical imperatives, a convergence between the socioeconomic systems of capitalism

and socialism (Kerr et al. 1960; Aron 1967). In the work of Burnham (1945) convergence, is based on a 'sectional' interpretation of management objectives. For Burnham, the managerial revolution is associated not so much with a shift from concentrated to dispersed private ownership, but from private to state ownership. This development is then associated with the emergence of a new dominant class of state enterprise managers exercising control in their own interests. Burnham's thesis has been criticized for its lack of clarity over who the managers in this revolution are, and whether it is in any sense reasonable to regard them as an homogeneous class (Dahrendorf 1959; Nichols 1969). The notion of a sectional managerialism, and the behavioural effects it may produce, however, has been taken up and developed separately by economists. Galbraith (1967) combines it with a view that bureaucratic organization is an inevitable feature of industrial life in the East and West, and that those in the organization mould the economic system as a whole to meet their own ends. He adopts an analysis of goal formation which rests heavily on the behaviouralist/organization theorists' view of the modern corporation as an adaptive organization (March and Simon 1958; Simon 1965; Cyert and March 1961). In his analysis, specialized technical and scientific managerial skills lead to the emergence within the large corporation of a loosely defined managerial 'technostructure'. Its members have a strong self-interested commitment to the survival of the organization to which they belong. This self-interest is pursued by adapting corporate goals towards planning the corporate environment, and by the pursuit of growth, and increases in size, subject only to a minimum profit constraint to keep the shareholders happy. This is the story made familiar in the earlier models of Baumol (1959) and Williamson (1964) and particularly in the distinctive dynamic work of Marris (1964).

In these more formal contributions, managers pursue higher salaries and corporate perks and generally attempt to divert corporate resources to serve their own interests. This usually boils down to aiming for greater output levels and faster growth than is consistent with maximizing the current stock market value of the corporation (taken as a direct proxy for stockholder welfare). The extent of management's discretion to do this depends upon a minimum profit constraint imposed by the capital market, or, in one version of Marris's original contribution upon sustaining a market value high enough to forestall a disciplinary takeover bid in the market for corporate control. In these sectional managerial models the essential shift in behaviour at the level of the individual firm is towards lower risk-taking (managers will avoid projects with highly variable profit streams, if downside risk threatens job security whilst superior profit performance has little impact on their remuneration package), and a faster rate of growth. The outcome at an aggregate level is not so clear, since individual corporations may grow by acquiring other existing corporations rather than by investing in new plant and equipment. It has been argued, however, that management discretion combined with institutional circumstances which restrict takeover possibilities has, in Japan for instance, contributed to a more dynamic overall aggregate growth performance (Odagiri 1981).

Opponents of the managerial capitalism thesis have argued that the separation of ownership from control has been over-emphasized or misinterpreted, that managerial discretion could not in any event exist to any significant degree, and that even if it did, it would not lead to significant changes in corporate behaviour since there is a congruence between the behavioural norms and the self-interest of major shareholders on the one hand, and top managers on the other.

The original Berle and Means thesis rested on particular historical, legal, and institutional developments in which the 'power vacuum' left by the emergence of the joint stock company and the dilution of ownership interests was filled by a salaried managerial class. Other outcomes are possible as circumstances differ. Thus, for example, starting from similar views about the emergence of a managerial group in charge of the day-to-day running of corporations, the 'finance capital' theorists have argued that the managers remain subordinate to the wishes of a small group of major shareholders, who retain interlocking

key positions on the boards of industrial companies, and on the boards of the financial institutions responsible for capital market intermediation. Although originally developed in a context in which equity capital markets were less important than the market in loan finance, the re-emergence of major shareownership groups in the form of pension funds and insurance companies (as noted by Berle himself: Berle 1960) and concentrated equity management groups, such as banks, has led to the revival of the theory as a rival to managerialist interpretations of contemporary capitalism (Hilferding 1910; Kotz 1978; Minns 1980; Scott 1986). Whilst the structural changes in capital markets on which this revival is based are clear enough, the extent to which they represent a subordination of managers to ownership or other interests, rather than a form of constraint, upon essentially, dominant 'inside' management is less obvious when account is taken of the way in which 'financial' directors are appointed to the boards of industrial companies, and their relative transience compared to insiders (Herman 1981).

Critiques based on the idea that managerial discretion is limited regard these issues as of second order importance anyway. Such critiques focus on conditions in the product and capital markets, and in the market for managers. They emphasize that the growth of concentrated market structures is the outcome of a competitive process by which the efficient come to account for a greater share of economic activity. In the absence of significant entry barriers no persistent monopoly power is possible. Even if it was, then the market for managers, and the stock market selection process, will together ensure that those in 'control' act so as to minimize costs. Thus, on this view of the world, the natural selection properties of the environment determine which policies have 'survival' value, and the discretionary role of managers is negligible (Alchian 1950; Friedman 1953; Becker 1962; Alchian and Kessel 1962; Manne 1965; Jensen and Ruback 1983; Jensen and Meckling 1976; Fama 1980; Demsetz 1974, 1983). There is sufficient empirical and theoretical analysis to suggest that these arguments cannot be wholly convincing. The evolutionary emergence of product market structures need not reflect dominant profit maximizing, as opposed to satisficing, decision rules (Winter 1964, 1971). The latter, in turn, may reflect the outcome of the activities of coalitions of interest groups internal to the large corporation, and thus re-admit managerial and other interpretations (Cyert and March 1963; Aoki 1984). Neither do the characteristics of the takeover mechanism as a key part of the stock market selection process, suggest that profit maximizing decision rules are superior to others such as growth maximization in avoiding raids by competing managements (Singh 1971, 1975; Mueller 1980). Finally, barriers to entry and the policies of dominant firms may lead to positions of market power, which, whilst not eternal, may be sufficiently persistent in particular cases to make it worthwhile to ask what use is made of any discretionary room for manoeuvre that is created (Scherer 1980; Mueller 1986).

As it happens, direct tests of the impact of patterns of ownership on corporate performance rarely allow for market power effects. Studies which do not control for this, report on average, small or negligible ownership impacts on profitability (e.g. Kania and McKean 1976; Monsen, Chiu, and Cooley 1968; McEachern 1975; Herman 1981). Studies which do allow for it, report mixed results. Thus Palmer (1973) reports manager-controlled companies earning lower profits than owner-controlled companies in markets with substantial market power, whilst Qualls (1976) reports negligible differences. Generally speaking, few studies of ownership impacts directly assess the extent to which such profit differences as do emerge are simply 'quiet life' effects as opposed to a conscious pursuit of alternative objectives such as growth, as hypothesized in the dynamic managerial models mentioned earlier. Multivariate tests, including both growth and profits, suggest an inferior performance in both dimensions for manager-controlled companies rather than a sacrifice of the latter in favour of the former (Holl 1975; Radice 1971). Whether these studies are identifying a supply of finance function rather than a demand for growth function remains, however, a moot point. As does the general assumption behind nearly all the empirical work in this area that 'control' categories can be

drawn up on the basis of identifying one or more minority shareownership groups defined at arbitrary levels of, say, 5% or 10% of total equity. The relative neglect of whether the ownership groups are personal or institutional; are located on-or off-board; are connected with founding family or other financial interests; and are transient or persistent; mean that they can only be a rough and ready guide to likely behavioural difference (Francis 1980a). Similarly, the neglect of the motivational impact of absolutely large, though in percentage terms, relatively small, equity holdings by managers is a further shortcoming. It reflects a general neglect of the underlying assumption of the managerialist models that owners, managers, and financiers have identifiably different objectives and behavioural norms which condition the objectives which corporations pursue when different interests dominate them (Francis 1980b; Baran and Sweezy 1968; Nichols 1969; Cosh and Hughes 1987).

It would be difficult, on the basis of the evidence so far, to make a strong case for, or against, the managerialist position that the separation of ownership from control has produced noticeable behavioural or performance differences between individual corporations with different degrees of separation. That the financial and organizational structure of the modern corporation has changed and is still evolving, is without dispute. The same cannot be said for the view that this has led to either a more soulful or socially responsible capitalism, or to a socio-economic structure in which a distinctly identifiable social group of 'managers' has come to exercise power in its own sectional interests.

See Also

- ▶ Administered prices
- ▶ Capitalism
- ▶ Corporate economy
- ▶ Entrepreneur
- ▶ Industrial organization
- ▶ Monopoly capitalism
- ▶ Multinational corporations
- ▶ Privatization

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58(June): 211–221.

Alchian, A.A., and Kessel, R.A. 1962. Competition, monopoly and the pursuit of pecuniary gain. In *Aspects of labour economics*, ed. National Bureau Committee for Economic Research. Princeton: Princeton University Press.

Aoki, M. 1984. *Cooperative game theory of the firm*. Oxford: Oxford University Press.

Aron, R. 1967. *The industrial society*. London: Weidenfeld & Nicolson.

Baran, P., and P. Sweezy. 1966. *Monopoly capital*. New York: Monthly Review Press; Harmondsworth: Penguin, 1968.

Baumol, W.J. 1959. *Business behaviour, value and growth*. New York: Macmillan.

Becker, G.S. 1962. Irrational behavior and economic theory. *Journal of Political Economy* 70(February): 1–13.

Bell, D. 1961. *The end of ideology*. New York: Collier, Macmillan.

Bell, D. 1974. *The coming of post-industrial society*. London: Heinemann.

Berle Jr., A.A. 1955. *The twentieth-century capitalist revolution*. London: Macmillan.

Berle Jr., A.A. 1960. *Power without property*. New York: Harcourt, Brace.

Berle Jr., A.A., and G.C. Means. 1932. *The modern corporation and private property*. New York: Commerce Clearing House.

Burnham, J.S. 1945. *The managerial revolution*. Harmondsworth: Penguin Books.

Cosh, A.D., and Hughes, A. 1987. *The anatomy of corporate control*. Department of Applied Economics, Cambridge University, Mimeo.

Cyert, R.J., and J.G. March. 1963. *A behavioral theory of the firm*. Englewood Cliffs: Prentice-Hall.

Dahrendorf, R. 1959. *Class and class conflict in an industrial society*. London: Routledge & Kegan Paul.

Demsetz, H. 1974. Two systems of belief about monopoly. In *Industrial concentration: The new learning*, ed. H.J. Goldschmid et al., 164–184. Boston: Little, Brown.

Demsetz, H. 1983. The structure of ownership and the theory of the firm. *Journal of Law and Economics* 26(2): 375–390.

Drucker, P.F. 1951. *The new society: The anatomy of the industrial order*. London: Heinemann.

Fama, E. 1980. Agency problems and the theory of the firm. *Journal of Political Economy* 88(April): 288–307.

Francis, A. 1980a. Families, firms and finance capital. *Sociology* 14(1): 1–27.

Francis, A. 1980b. Company objectives, managerial motivation and the behaviour of large firms: An empirical test of the theory of managerial capitalism. *Cambridge Journal of Economics* 4(4): 349–361.

Friedman, M. 1953. *Essays in positive economics*. Chicago: University of Chicago Press.
Galbraith, J.K. 1967. *The new industrial state*. London: Hamish Hamilton.
Herman, E.S. 1981. *Corporate control, corporate power*. Cambridge: Cambridge University Press.
Hilferding, R. 1910. *Finance capital*. London: Routledge & Kegan Paul, 1981.
Holl, P.J. 1975. Effect of control type on the performance of the firm in the UK. *Journal of Industrial Economics* 23(4): 257–271.
Jensen, M.C., and W. Meckling. 1976. Theory of the firm: Managerial behaviour, agency costs and ownership structure. *Journal of Financial Economics* 3(October): 305–360.
Jensen, M.C., and R. Ruback. 1983. The market for corporate control: The scientific evidence. *Journal of Financial Economics* 11(April): 5–50.
Kania, J.J., and J.R. McKean. 1976. Ownership, control and the contemporary corporation; A general behavioural analysis. *Kyklos* 29: 272–291.
Kerr, C., et al. 1960. *Industrialism and industrial man*. Berkeley: University of California Press.
Kotz, D.M. 1978. *Bank control of large corporations in the United States*. Berkeley: University of California Press.
Manne, H.G. 1965. Mergers and the market for corporate control. *Journal of Political Economy* 73(April): 110–120.
March, J.G., and H.A. Simon. 1958. *Organizations*. New York: Wiley.
Marris, R.L. 1964. *The economic theory of 'managerial' capitalism*. London: Macmillan.
Mason, E.S. (ed.). 1960. *The corporation in modern society*. Cambridge, MA: Harvard University Press.
Mayo, E. 1949. *The social problems of an industrial civilization*. London: Routledge & Kegan Paul.
McEachern, W.A. 1975. *Managerial control and performance*. Lexington: D.C. Heaton.
Minns, R. 1980. *Pension funds and British capitalism*. London: Heinemann.
Monsen, R.J., J.S. Chiu, and E.D. Cooley. 1968. The effects of separation of ownership and control on the performance of the large firm. *Quarterly Journal of Economics* 82(August): 435.
Mueller, D.C. (ed.). 1980. *The determinants and effects of mergers*. Cambridge, MA: Oelgeschlager, Gunn and Hain.
Mueller, D.C. 1986. *Profits in the long run*. Cambridge: Cambridge University Press.
Nichols, T. 1969. *Ownership, control and ideology*. London: George Allen & Unwin.
Odagiri, H. 1981. *The theory of growth in a corporate economy*. Cambridge: Cambridge University Press.
Palmer, J.P. 1973. The profit-performance effects of the separation of ownership from control in large US industrial corporations. *Bell Journal of Economics and Management Science* 4(1): 293–303.
Qualls, P.D. 1976. Market structure and managerial behaviour. In *Essays on industrial organization in honour of Joe S. Bain*, ed. R. Masson and P.D. Qualls. Cambridge, MA: Ballinger.
Radice, H.K. 1971. Control type profitability and growth in large firms. *Economic Journal* 81(September): 547–562.
Scherer, F.M. 1980. *Industrial market structure and economic performance*. Chicago: Rand-McNally.
Scott, J. 1986. *Capitalist property and financial power, a comparative study of Britain, the United States and Japan*. Brighton: Wheatsheaf Books.
Singh, A. 1971. *Takeovers: Their relevance to the stock market and the theory of the firm*. Cambridge: Cambridge University Press.
Singh, A. 1975. Takeovers, economic natural selection and the theory of the firm: Evidence from the United Kingdom experience. *Economic Journal* 85(September): 497–515.
Veblen, T. 1924. *Absentee ownership and business enterprise in recent times*. London: George Allen & Unwin.
Williamson, O.E. 1964. *The economics of discretionary behavior*. Englewood Cliffs: Prentice-Hall.
Winter, S.G. 1964. Economic 'natural selection' and the theory of the firm. *Yale Economic Essays* 4(1): 225–272.
Winter, S.G. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85(2): 237–261.

Manchester School

William D. Grampp

Keywords

Anti-Corn Law League; Bright, J.; Cobden, R.; Corn Laws; Free trade; Manchester School; Navigation Laws; Philosophic radicalism

JEL Classifications

B1

The Manchester School was the name given by Disraeli after the event to the leaders of the successful agitation conducted between 1838 and 1846 to abolish the Corn Laws. It is wrongly associated with the arch-advocacy of laissez-faire. The people of the School were not in fact united by any single idea, other than believing in the complete and immediate repeal of the tariff on grain.

Within the School there were five discernible groups in the sense of there being five different reasons why people wanted repeal or purposes that directed them.

Some were compatible with others, and one group could agree with another over what was important but differ over how important it was. The arguments that each group made do not, when taken together, constitute a cogent or even coherent whole but taken separately could be both, and are always interesting. Moreover, the campaign for repeal is itself an instructive event in the history of economic policy.

One group was the mill-owners of Lancashire who provided most of the money for the campaign and formed the National Anti-Corn Law League to conduct it. Some believed that repeal, by reducing the price of bread, would reduce money wages, hence the cost of production in their mills. The belief comes from the Ricardian principle that real wages are constant in the long run. It could have made the businessmen believe the export of grain should be protected, since that too could reduce its price, hence have placed them in the interesting but not unusual position of half-believing in the free market.

They in fact did not support protection because a greater reason for their wanting repeal was to increase the export of manufactured goods. The economic argument most often made was that importing more grain would provide foreigners with more income to spend on British exports, with the result that income and employment would increase at home. The mill-owners were repeatedly accused of simply wanting to cut wages. Cobden privately warned them to stay out of the repeal campaign if they could not come in with clean hands. Publicly he offered to support a Factory Bill of Lord Ashley – the 'universal syllabub of philanthropic twaddle', in Carlyle's description – if Ashley would pay his farm hands what the workers in Cobden's factory were paid. The offer was declined.

Another economic augument for repeal was that it would retard the growth of manufacturing abroad and so keep Britain in its leading industrial position. Why the owner of a small mill would profit by his country's having more mills than any other was not made clear (although he might by way of an externality of some sort). The argument is nevertheless noteworthy. It was revived after 1945, when the undeveloped countries hastened to industrialize in the belief that doing so was a necessary condition of their progress. The argument is also part of the curious notion, entertained by historians, that Britain's free trade was an instrument of its imperialism. They reason that Britain, by keeping others in a non-industrial condition, could dominate them, exploit them, and/or make them dependent on it.

Why one country would choose to be mistreated by another when it could choose a trading partner that did not mistreat it, as in a system of free trade it could do, is not explained. Or is there an explanation of why a dollar's worth of manufactured goods adds more to total welfare than a dollar's worth of goods that are not manufactured?

Among the businessmen working for repeal were those who believed it would make life better for the lower classes. They have been called the humanitarian employers. They did more for their workers than the market or the law required, providing schools for the children, reading rooms and meeting places for the men and women, helping them to form friendly societies, cooperatives and cultural groups. Some employed a 'salaried visitor' (social worker) to call at the homes. These business people also undertook to improve the communities where they were established. One such effort was the Manchester Statistical Society which collected information on living conditions and used it to improve them. The Greg family stood out in this group.

The radical businessmen, working on a larger scale than the humanitarians, aspired to improve the nation and the world. In economic affairs their great end was free trade and after the repeal of the Corn Laws they had a part in the abolition of the Navigation Laws. In politics they looked toward democratic government and worked to extend the franchise until all adult males had the right to vote. The radicals believed free trade would first increase the influence of the business classes, increase their members in Parliament, then (in a way not fully explained) increase the power of the working classes.

John Bright was the leader of this group, which itself was the Manchester version of the middle-class radicalism of the time. It had a finger or a hand or more in most reform movements, great and small, from the abolition of slavery and removal of religious disabilities to the penny post and repeal of the taxes on knowledge. The radicals were disrespectful of authority, indifferent to custom, unmindful of the ridiculous figure they often cut, and they were meddlesome, tiresome, persistent and effective. Like Pancks, what they did, they did, they did indeed, and when they finished there were noble institutions in ruins.

The Philosophic or London Radicals had a different place from that of the radical businessmen, grounding their reform on a considered application of Bentham's utilitarianism and conducting themselves in the mannerly, measured way that made them heard and respected but unheeded and ineffectual. They did not care for the rough and ready way of Manchester and had to be reminded of where they were before it took on the repeal of the Corn Laws. Before them, Charles Villiers, a leading Benthamite, had each year moved in the House that it constitute itself a committee of the whole to consider the repeal of the Corn Laws, and each year the motion was defeated. The leadership of the free trade bloc passed to Cobden when he became a Member of Parliament, an instance, his friends said, of talent giving way to genius. Francis Place, who was on the edge of the London Radicals, put things plainly and said that when the Manchester people wanted something done they did it.

Cobden represented the pacifists of the School. They believed that trading nations had a material interest in peace, an idea Ricardo had stated in his *Essay on Profits* in 1815, and that they were natural friends by virtue of meeting on the market, an idea Ricardo was too realistic to entertain. Oddly, the pacifists seem not to have noticed they could have drawn an argument from the *Wealth of Nations*. No pacifist himself, Smith said Britain should not engage in trade that would diminish its military power. The implication is that free trade makes nations unable to go to war as well as unwilling.

The pacifists, although not the largest group within the School, were even more influential than the radicals. Cobden wanted to graft the peace movement onto the repeal campaign although he would not permit the franchise to be so joined, as Bright wanted to do. After repeal, the franchise had more public support than the peace movement and grew until all adults had the vote. Nevertheless those who believe free trade is conducive to peace can and do point out that the 19th century was a time when trade was freer than ever and was the only century in recent history when there has not been a world war.

Cobden wanted free trade because it would bring peace, Bright because it would bring the franchise. Others in the School had each of them his own purpose. They made common cause for seven years until the Corn Laws were brought down, then returned to their separate ways.

The Manchester School was a coalition around a single issue. It was not a group of ideologues committed to laissez-faire, as historians have carelessly said, nor did it express the pure spirit of the middle class, as some contemporaries believed. It was not a rent-seeking force, as Public Choice economists are tempted to say, nor did it preach the principles of huckstering (Disraeli), nor were its leaders 'bartering Jews' (Engels), nor were they 'the official representatives of the bourgeosie' (Marx). If the Manchester School is to be described simply, it was a remarkably successful effort to remove a major obstacle in the way of the market.

See Also

- ▶ Bright, John (1811–89)
- ▶ Cobden, Richard (1804–1865)

References

Grampp, W.D. 1960. *The Manchester school of economics*. Stanford: Stanford University Press. London: Oxford University Press.

Hirst, F.W., ed. 1903. *Free trade and other fundamental doctrines of the Manchester School*. London: Harper & Bros.

McCord, N. 1958. *The anti-corn law league 1838–1846*. London: Allen & Unwin.
Morley, J. 1881. *The life of Richard Cobden*. London: Chapman & Hall.
Prentice, A. 1853. *History of the anti-corn law league*. London: W. & F.G. Cash. *Proceedings of the Chamber of Commerce and Manufactures at Manchester. 1821–1865*. Mss. at the Manchester Central Library.
Students in the Honours School of History in the University of Manchester and Arthur Redford. 1934. *Manchester merchants and foreign trade 1794–1858*. Manchester: Publications of the University of Manchester, Economic History Series.
Trevelyan, G.M. 1913. *The life of John Bright*. London: Constable.

Mandated Employer Provision of Employee Benefits

Jonathan Gruber

Abstract

Mandated employer provision of social benefits is of rising importance in the United States. As highlighted by Summers (1989), the efficiency losses from such mandates may be much lower than those of taxation due to tax–benefit linkages. I review the theory underlying this observation and the empirical evidence which documents full shifting to wages (and therefore little efficiency cost) of mandated benefits. A host of important questions about mechanisms remains unanswered, however.

Keywords

Adverse selection; Anti-discrimination law; Cost shifting; Health insurance; Labour supply; Mandating employer provision of employee benefits; Minimum wages; Payroll taxation; Unemployment insurance; Workers' compensation insurance

JEL Classifications

H2

The provision of social benefits can be financed in a number of different ways: through broad income taxation, through taxation of payroll only, or through mandates on employers to provide those benefits for their employees. The last channel is one of sizable and growing importance in the United States, although less so in other nations that tend to rely more on tax-financed government provision. Yet, until the late 1980s, the impacts of mandates were not much studied. The implicit assumption in economic analysis was that such mandates could be analysed using the standard tools of tax incidence and efficiency.

A very influential article by Summers (1989) changed all that. Summers pointed out that mandating employer provision of benefits to their employees had two effects on labour market equilibrium. On the one hand, a reduction in labour demand naturally accompanies the imposition of extra costs on employers. On the other hand, however, mandates should also cause an outward shift in labour supply, since individuals are now being effectively compensated more highly for their labour; they are receiving their previous compensation plus the mandated benefit. This shifts more of the costs of benefits to workers and reduces the deadweight loss from their provision. Indeed, as Summers pointed out, if employees valued the mandated benefit at its cost to the employer, then these supply and demand shifts would be equal. The end result would be 'full shifting to wages': a decline in wages by exactly the cost of the benefit with no impact on total labour supplied to the market and no efficiency consequences. Employees would simply be buying a benefit they value with their wages.

This article inspired a large follow-up literature, mostly empirical, investigating the equity and efficiency properties of mandates. I review that literature here, in three steps. First, I comment on the theoretical points made by Summers. Second, I discuss the empirical evidence available on the impacts of mandates. Finally, I discuss the key unanswered questions that must be addressed by future research.

Theoretical Background

Summers' analysis was as straightforward as it was insightful, highlighting the impacts of mandates in a simple demand and supply framework. The mathematics behind this analysis is explored in Gruber (1992), Gruber and Krueger (1991) and Anderson and Meyer (1997). These analyses show that the incidence of mandated benefits depends on the elasticities of supply and demand, as with any tax, along with a new parameter: the valuation of the benefit by employees. If valuation is equal to the cost paid by the employer for the benefit, then there is full shifting to wages.

But this analysis misses an important point: Summers' analysis is in no way restricted to mandates. Indeed, the analysis is exactly the same for Unemployment Insurance, a US programme which provides tax-financed benefit to unemployed workers. The key to Summers' analysis is not the form of provision (mandate or tax); rather, the key is that the benefits are *restricted to workers*, generating the labour supply increase that offsets some of the efficiency consequences of the intervention. For example, a payroll tax-financed expansion of health insurance to workers fits into this framework, but a payroll tax-financed expansion of health insurance to all individuals in society does not. In the latter case, there would not be the corresponding increase in labour supply, since individuals would not have to work to receive the benefit.

Another question raised by Summers' analysis is this: if there were full incidence on wages, why wouldn't employers simply provide the benefit voluntarily? Why is government coercion necessary to promote employer provision of a benefit fully valued by employees? The best answer here, as pointed out by Summers, is that there may be market failures that lead employers to not reflect workers' valuation of this programme without a government mandate. Most obviously, adverse selection in the market for benefits could cause employer reluctance to be, for example, the one employer in town that offered health insurance or paid maternity leave. This standard adverse selection problem may keep employers from offering benefits that are fully valued by employees. (Indeed, if there is such a market failure, it is feasible that a programme such as workers' compensation could *raise* the quantity of labour in the market. If workers value workers' compensation at more than its cost to employers – as might be the case if workers are risk averse – the labour supply curve would shift out by more than the demand curve shifted in; workers would be willing to accept a wage cut of *more* than the cost of workers' compensation in order to have this benefit. This would actually raise employment.)

Empirical Evidence

During the 1990s a large number of articles explored the empirical impact of mandates, in particular the extent to which mandated costs were shifted to wages. This literature is reviewed in detail in Gruber (2001); I provide an overview here. The consensus of this literature is that, over the medium to long term, the cost of mandates is fully reflected in wages.

Gruber and Krueger (1991) provide the first such analysis, dealing with increases in the employer costs of Workers' Compensation (WC) insurance across US industries and states over time. WC provides cash benefits and health coverage to workers injured on the job, and much of the variation in costs in the authors' data comes from increases in the health care component of this programme. They focus on workers in five industries for which WC costs are high and rapidly growing; in some industries and states, these costs amounted to over 25 per cent of payroll by 1987, the end of their sample period. They use both micro-data on wages and aggregate data on employment and wages by state/industry. They include state and industry fixed effects in their models, so that they are controlling for general differences in pay across industries and places, and estimating only how that pay changed when the costs of WC rose. In both data-sets, they find that for these sets of industries 85 per cent of increases in workers compensation costs were shifted to wages.

Anderson and Meyer (1997) undertake a similar analysis for Unemployment Insurance (UI),

which provides cash benefits to unemployed workers. This programme is not a mandate, but it should operate in the same fashion as it levies payroll taxes on firms to provide benefits to their workers. Anderson and Meyer's conclusion is similar to Gruber and Krueger: general differences in UI payroll taxes appear to be fully reflected in wages with little effect on labour supply.

There is also a long literature on the impact of payroll taxation on wages that is reviewed by Hamermesh (1987). This literature is much more mixed in its conclusions, although the variation in payroll taxes mostly comes over time, and it is difficult to estimate its incidence separately from other time series factors in the United States where there is little variation across workers in payroll tax rates. More recent evidence is consistent, however, with the notion of full shifting to wages in other countries (for example, Gruber 1997).

Labour supply is not simply a discrete choice, however, but rather a combination of participation and hours of work decisions. Increases in costs will effect both the supply of and the demand for work hours conditional on participation. From the employer perspective, increases in health insurance costs are an increase in the fixed cost of employment and are as a result more costly (as a fraction of labour payments) for low-hours employees. Employers will therefore desire increased hours by fewer workers, lowering the cost per hour of the health insurance for a given total labour supply. Of course, if the wage offset is lower for low-hours workers, workers will demand the opposite outcome: there will be increasing demand for part-time work, with hours falling and employment increasing. Moreover, since part-time workers may be more readily excluded from health insurance coverage, there may also be a countervailing effect on the employer side, as full-time employees are replaced with their (uninsured) part-time counterparts. In this case as well, hours would fall and employment would rise. Thus, the effect on hours of work is uncertain. Several studies have addressed this issue, and the general consensus is that mandating fixed costs of employment leads to rises in hours and falls in employment (Gruber 1994; Cutler and Madrian 1998).

Remaining Questions

While there have been significant gains in our understanding of the impacts of mandates, important questions remain unanswered. The most important is the question of heterogeneity across workers: how do mandates affect workers differentially within the workplace? Consider the example of mandated health insurance. The cost of health insurance will not be uniform throughout the workplace; costs are higher for family insurance than for individual coverage, or for older workers than for younger workers. In the limit, with extensive experience rating, costs vary worker by worker, depending on their underlying health status. Gruber (1992) extends the model of Gruber and Krueger (1991) to the case of two groups of workers, where costs increase for one but not the other. If there is group-specific shifting, then the solution collapses to the one group model. If not, however, the substitutability of these groups will also determine the resulting labour market equilibrium; in general, there will be effects on both the group for which costs increase and the group for which they do not.

In practice, there may be a number of barriers to group-specific, and in particular individual-specific, shifting. Most obviously, there are anti-discrimination regulations which prohibit differential pay for the same job across particular demographic groups, or which prevent differential promotion decisions by demographic characteristic. Workplace norms which prohibit different pay across groups or union rules about equality of pay may have similar effects. Thus, a central question for incidence analysis is *how finely* firms can shift increased costs to workers' wages. If there is imperfect group- or worker-specific shifting, there may be pressure on employers to discriminate against costly workers in their hiring decisions.

Two studies suggest that there is within-workplace shifting to wages. Gruber (1994) studied the effect of state laws (and a follow-up federal law) that mandated in the mid-1970s that the costs of pregnancy and childbirth be covered comprehensively. Before this time, health insurance plans provided very little coverage for the costs associated with normal pregnancy and childbirth, while

providing generous coverage for other medical conditions. This distinction was viewed as discriminatory by some state governments, leading to the state laws mandating that pregnancy costs be covered as completely as other medical costs. These laws significantly increased the insurance costs for women of childbearing age in those states, thereby raising the costs of employing a specific group of workers (or their husbands, who may provide them with insurance). I estimated full shifting to wages for these groups. Further corroborating evidence on this point is provided by Sheiner (1999), who found that, when healthcare costs rise in a city, the wages of workers who have the highest costs (older and married workers) fall the most.

This research suggests that within-workplace shifting to wages is possible. The news here is good for efficiency: mandates which have differential effects across broad groups of workers will not necessarily lead to displacement of the high-cost group. The news is potentially bad for equity, however: other groups will not crosssubsidize the high costs imposed on one group in the workplace. In any case, neither of these studies addresses the extent to which within-firm shifting is possible; in the limit, it is hard to conceive that employers could shift costs to wages on a worker-by-worker basis.

Other questions have not been addressed at all by the literature. First, how rapidly does shifting to wages occur? Despite the evidence that mandates are fully reflected in wages, employers vociferously oppose mandated benefits as costing jobs. The reason for their opposition could be that wages cannot adjust quickly enough in the short run to offset displacement effects; the studies cited earlier show full shifting only over several-year periods.

Second, what are the effects of existing constraints on compensation design in the labour market? For example, for workers already at the minimum wage, firms will be unable to shift to wages increases in the cost of health insurance. Similarly, union contract or other workplace pay norms may interfere with the adjustment of wages to reflect higher costs. These institutional features could increase the disemployment effects of rising health costs.

Third, what is the underlying structural mechanism behind a finding of full shifting to wages? In the simple labour market framework above, there are two reasons why increased costs might be shifted to wages: because individuals value the benefits that they are getting fully; or because labour supply is perfectly inelastic. Disentangling these alternatives is very important for future policy analysis. Consider the example of national health insurance, which is financed by a mandate, with an additional payroll tax to cover non-workers. If the full shifting documented earlier is due to full employee valuation with somewhat elastic labour supply, then national health insurance will have important disemployment effects, since labour supply will not increase in response to a benefit that is not restricted to workers. If full shifting is due to inelastic supply, however, then the population which is receiving benefits is irrelevant; in any case the costs will be passed on to workers' wages, so national health insurance will not cause disemployment. Existing evidence, as reviewed in Gruber (2001), is mixed on which of these channels is at work.

Conclusion

Since the early 1990s there has been a substantial growth in research on mandated benefits. The conclusions from the work to date are clear: the costs of mandated benefits are fully shifted onto wages, with little impact on total labour supply. But important questions about the mechanisms behind such shifting remain unanswered.

Bibliography

Anderson, P., and B. Meyer. 1997. The effect of firm specific taxes and government mandates with an application to the U.S. Unemployment Insurance program. *Journal of Public Economics* 65: 119–144.

Cutler, D., and B. Madrian. 1998. Labor market responses to rising health insurance costs: evidence on hours worked. *RAND Journal of Economics* 29: 509–530.

Gruber, J. 1992. *The efficiency of a group-specific mandated benefit: Evidence from health insurance benefits for maternity*. Working paper no. 4157. Cambridge, MA: NBER.

Gruber, J. 1994. The incidence of mandated maternity benefits. *American Economic Review* 84: 622–641.
Gruber, J. 1997. The incidence of payroll taxation: evidence from Chile. *Journal of Labor Economics* 15(3, Part 2): S72–S101.
Gruber, J. 2001. Health insurance and the labor market. In *The handbook of health economics*, ed. J. Newhouse and A. Culyer. Amsterdam: North-Holland.
Gruber, J., and A. Krueger. 1991. The incidence of mandated employer-provided insurance: Lessons from Workers' Compensation insurance. In *Tax policy and the economy 5*, ed. D. Bradford. Cambridge, MA: MIT Press.
Hamermesh, D.S. 1987. Payroll taxes. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London: Macmillan.
Sheiner, L. 1999. *Health care costs, wages, and aging. Finance and economics discussion series 1999–19*. Washington, DC: Board of Governors of the Federal Reserve System.
Summers, L. 1989. Some simple economics of mandated benefits. *American Economic Review* 79: 177–183.

Mandel, Ernest (1923–1995)

Geoffrey M. Hodgson

Keywords

Capitalism; Central planning; Kondratieff cycles; Labour theory of value; Long waves; Mandel, E; Marx, K; Neo-liberalism; Socialism; Soviet Union; Trotsky, L; Unemployment

JEL Classifications

B31

Ernest Mandel was born of Jewish parents in Frankfurt-am-Main on 5 April 1923. The family emigrated to Antwerp. By 1939 he was actively involved in socialist and trade union politics. When the Nazis invaded Belgium in 1940, he became a member of the resistance. On three occasions he was arrested and imprisoned, but each time he escaped. He was arrested for a final time in October 1944 and liberated by the Allies in March 1945. He obtained a higher education in Brussels and Paris. His name was prominent in academia in the 1960s and 1970s when Marxism and Trotskyism enjoyed significant popularity, particularly among university students. He died on 20 July 1995.

His *Marxist Economic Theory* was first published in French in 1962 and in English in 1968. When student revolts and labour unrest broke out in the late 1960s, Mandel's text and the much shorter *Introduction to Marxist Economic Theory* (1967) were available for the growing numbers interested in Marxist economics. His *Marxist Economic Theory* was widely praised and his *Introduction* sold over half a million copies and was translated into 30 languages.

His *Formation of the Economic Thought of Karl Marx* was published in French in 1967 and in English in 1971. It was one of the first works in English to analyse Marx's *Grundrisse,* which did not appear in complete form in English until 1973.

In his *Europe* vs. *America* – published in German in 1968 and in English in 1970 – he predicted relative economic decline and increasing 'public squalor' in the United States, sustained rapid economic growth in Japan, and the achievement of productivity levels in the western European 'core' regions to rival those in America.

In his *Late Capitalism* – published in German in 1972 and in English in 1975 – he revisited the idea that capitalism was subject to repeated waves of boom and stagnation in 45–60 year cycles. Not only did Mandel predict the downturn of the 1970s on the basis of this analysis, but also this work help to revive academic interest in the study of long waves, which has continued to the present day. However, his analysis has been criticized for misunderstanding Trotsky's criticisms of Kondratiev (Day 1976) and lacking a plausible mechanism to explain the complete long-wave cycle (Tylecote 1992).

Mandel wrote introductions to the new English translations of the three volumes of Marx's *Capital,* published by Penguin (Marx 1976, 1978, 1981). He was obliged to consider the stormy technical debates in the 1970s over the labour theory of value and Marx's theory of the tendency of the rate of profit to fall (Steedman 1977). However, instead of addressing the detailed critical arguments, he simply brushed them aside.

In 1978 Mandel was invited by the University of Cambridge to give the prestigious Alfred Marshall Memorial lectures. These were published as *Long Waves of Capitalist Development* (1980): a restatement and development of ideas in *Late Capitalism.* Further weaknesses in his position emerged when it became clear that mass unemployment in the West was not leading to political advances for socialism. Instead the period saw a resurgent political individualism and neoliberalism.

Like Trotsky, Mandel opposed the view that the Soviet-type economies were another type of 'capitalism', envisaged a collapse of the Soviet regimes, and expected that the working class would rise up in defence of state planning and nationalized property. Even after their collapse in 1989–91, in his *Power and Money* (1992) he hoped for a new workers' movement in eastern Europe and predicted that capitalism would not be easily re-established. Overall, the theoretical weakness of his outlook became increasingly clear in the last 15 years of his life.

See Also

- ▶ Capitalism
- ▶ Kondratieff Cycles
- ▶ Marx's Analysis of Capitalist Production
- ▶ Socialism
- ▶ Socialism (New Perspectives)
- ▶ Soviet Economic Reform
- ▶ Trotsky, Lev Davidovitch (1879–1940)

Selected Works

1967. *An introduction to Marxist economic theory.* New York: USA.
1968. *Marxist economic theory,* 2 vols. Trans. B. Pearce from the French edition of 1962. London: Merlin.
1970. *Europe* vs. *America: Contradictions of imperialism.* Trans. from the German edition of 1968. New York: Modern Reader.
1971. *The formation of the economic thought of Karl Marx: 1843 to capital.* Trans. B. Pearce from the French edition of 1967. London: NLB.
1975. *Late capitalism.* Trans. from the German edition of 1972. London: NLB.
1980. *Long waves of capitalist development: The Marxist interpretation.* Cambridge: Cambridge University Press.
1992. *Power and money: A marxist theory of bureaucracy.* London: Verso.

Bibliography

Day, R. 1976. The theory of long waves: Kondratiev, Trotsky, Mandel. *New Left Review* 99: 67–82.
Marx, K. 1976. *Capital.* Vol. 1. Trans. B. Fowkes from the fourth German edition of 1890. Harmondsworth: Pelican.
Marx, K. 1978. *Capital.* Vol. 2. Trans. D. Fernbach from the German edition of 1893. Harmondsworth: Pelican.
Marx, K. 1981. *Capital.* Vol. 3. Trans. D. Fernbach from the German edition of 1894. Harmondsworth: Pelican.
Steedman, I. 1977. *Marx after Sraffa*. London: NLB.
Tylecote, A. 1992. *Long waves in the world economy: The present crisis in historical perspective*. London: Routledge.

Mandeville, Bernard (1670–1733)

N. Rosenberg

Keywords
Division of labour; Laissez-faire; Mandeville, B.; Self-interest; Smith, A

JEL Classifications
B31

Mandeville was born in or near Rotterdam in 1670 and died in Hackney, London, in 1733. He was awarded the degree of Doctor of Medicine from the University of Leyden in 1961. He took up the practice of medicine, specializing in the 'Hypochondriack and Hysterick Diseases', a subject on which he later published a treatise. Mandeville travelled to England, married there in 1699, and lived in England for the rest of his life. He was very widely read in the 18th century. His writings

have often led to his being referred to as a satirist, but that is an inadequate and misleading classification.

Although Mandeville was not an economist, his writings were influential in shaping the direction of economic thinking in the 18th century. In 1705 he published a pamphlet, in doggerel verse, under the title *The Grumbling Hive: Or Knaves turn'd Honest.* In 1714 it was republished under its better-known title, *The Fable of the Bees: or, Private Vices, Publick Benefits.* This and subsequent editions included extensive expansions, clarifications and 'vindications' of his earlier themes. The grumbling hive was originally a thriving and powerful community. When, however, its inhabitants were suddenly and miraculously converted from a vicious to a virtuous moral condition, the community was swiftly reduced to an impoverished and depopulated state.

Mandeville's central theme is that public benefits are the product of private vices and not of private virtues. His paradox, which was widely regarded as scandalous, was achieved by employing a highly ascetic and self-denying definition of virtue. Since behaviour that could be shown to be actuated by even the slightest degree of self-regarding motive – pride, vanity, avarice or lust – was classified as vice, Mandeville had little difficulty in concluding that a successful social order must inevitably be one where public benefits are built upon a foundation of private vices.

> . . . I flatter myself to have demonstrated that, neither the Friendly Qualities and kind Affections that are natural to Man, nor the real Virtues he is capable of acquiring by Reason and Self-Denial, are the Foundation of Society; but that what we call Evil in this World, Moral as well as Natural, is the grand Principle that makes us sociable Creatures, the solid Basis, the Life and Support of all Trades and Employments without Exception: That there we must look for the true Origin of all Arts and Sciences, and that the Moment Evil ceases, the Society must be spoiled, if not totally dissolved. (Mandeville 1732, vol. 1, p. 369)

What was of more enduring significance in Mandeville's views was his forceful and unapologetic popularization of the belief that socially desirable consequences would flow from the individual pursuit of self-interest. It is an essential part of Mandeville's argument that a viable social order can emerge out of the spontaneous actions of purely egoistic impulses, requiring neither the regulation of government officials, on the one hand, nor altruistic individual behaviour, on the other.

> As it is Folly to set up Trades that are not wanted, so what is next to it is to increase in any one Trade the Numbers beyond what are required. As things are managed with us, it would be preposterous to have as many Brewers as there are Bakers, or as many Woolen-drapers as there are Shoe-makers. This Proportion as to Numbers in every Trade finds it self, and is never better kept than when nobody meddles or interferes with it. (Mandeville 1732, vol. 1, pp. 299–300)

Thus, Mandeville enunciates a vision of an economy that organizes itself and that allocates resources through the market place. Although there is no serious analysis of the workings of the market mechanism, there is the clear assertion that the unregulated market provides a system of signals and inducements such that the interactions of purely egoistic motives will somehow produce results that will advance the public good.

In developing his views, Mandeville offered many acute observations on the causes as well as the consequences of the division of labour in society. He regarded the division of labour as the great engine of economic improvement over the ages. It is the most reliable way for 'savage People' to go about 'meliorating their Condition'. For

> . . . if one will wholly apply himself to the making of Bows and Arrows, whilst another provides Food, a third builds Huts, a fourth makes Garments, and a fifth Utensils, they not only become useful to one another, but the Callings and Employments themselves will in the same Number of Years receive much greater Improvements, than if all had been promiscuously follow'd by every one of the Five. (Mandeville 1729, vol. 2, p. 284)

Although one can identify a number of possible precursors to Adam Smith's celebrated views on the division of labour, it is well established that Smith had in fact read and digested Mandeville carefully. Smith's marvellous description of the extensive division of labour involved in the production of a day-labourer's coat, with which he

closes the first chapter of the *Wealth of Nations,* may be traced to Mandeville's earlier treatment of the same subject – a treatment which, indeed, Smith extensively paraphrases. Moreover, the passage in the *Wealth of Nations* containing the often quoted statement that 'It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest' (Smith 1776, p. 14) is in a direct lineage from Mandeville's earlier observation:

> . . . The whole Superstructure [of Civil Society] is made up of the reciprocal Services, which Men do to each other. How to get these Services perform'd by others, when we have Occasion for them, is the grand and almost constant Sollicitude in Life of every individual Person. To expect, that others should serve us for nothing, is unreasonable; therefore all Commerce, that Men can have together, must be a continual bartering of one thing for another. (Mandeville 1729, vol. 2, p. 349)

Thus Mandeville was, in some important respects, an early advocate of laissez-faire (although this advocacy did not extend to foreign trade, where Mandeville's views were still distinctly Mercantilist). He articulated a vision of the role of the division of labour in society, and of the forces making for social change and evolution, as well as for social cohesion, that were in many respects distinctly precocious, and that exercised a powerful influence in shaping the intellectual agenda of economists and other social scientists later in the 18th century.

Selected Works

1714; 1729; 1732. *The fable of the bees.* 2 vols, ed. F.B. Kaye. London: Oxford University Press, 1924. (This is the definitive edition.)

Bibliography

Hayek, F.A. 1978. Dr. Bernard Mandeville. In *New studies in philosophy, politics, economics and the history of ideas*, ed. F.A. Hayek. Chicago: University of Chicago Press.

Primer, I., ed. 1975. *Mandeville studies*. The Hague: Martinus Nijhoff.

Rosenberg, N. 1963. Mandeville and laissez-faire. *Journal of the History of Ideas* 24 (2): 183–196.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 1937. New York: Modern Library.

Stephen, L. 1876. *History of English thought in the eighteenth century*. 2 vols. London: Smith, Elder.

Viner, J. 1953. *Introduction to Bernard Mandeville, a letter to Dion*. Los Angeles: Augustan Reprint Society, Publication No. 41.

Mangoldt, Hans Karl Emil von (1824–1868)

H. C. Recktenwald

Keywords

Entrepreneurship; Equilibrium; Historical School, German; Joint supply and demand; Mangoldt, H. K. E. von; Multiple equilibria; Price theory; Product life cycle; Profit and profit theory; Rents of differential ability

JEL Classifications

B31

Mangoldt was born in Dresden in 1824 and died in 1868 of a heart attack after a short life. He was an eminent theorist in economics, yet greatly underrated by his German contemporaries. He received his doctorate in Tübingen (1847) and was afterwards a civil servant in the Ministry of Foreign Affairs – a post he resigned for political reasons – and became editor of the official *Weimarer Zeitung* (1852). His academic career began in 1855 as Privatdozent in Göttingen and ended as Professor of Political Science and Political Economic after only six years (1862–8) at the University of Freiburg (Breisgau).

Mangoldt ranks among those pioneers in Germany, like von Thünen, von Buquoy, von Hermann, Gossen and Launhardt, who applied formal analysis to explain economic phenomena.

Yet the predominant influence of the Historical School diminished the impact of his methods and ideas on German university economists. He shared this fate with Cournot and Walras. A second reason for this underrating of his pioneering achievements at home was his strong interest in classical economics. Thus it is not surprising that his reputation was much higher in England (via Edgeworth and Marshall) than in 19th-century Germany.

In his most important books, *Unternehmergewinn* (1855) and *Grundriss* (1863), he argues in the classic tradition, examines its hypotheses in the light of economic and political reality and modifies them considerably. Like Cournot (earlier) and Marshall (later) Mangoldt uses a novel apparatus of partial analysis – Frisch's microanalysis –to expound originally a mathematical theory of prices that goes far beyond Cournot. He describes in a very modern way the process from one equilibrium to another, analyses multiple equilibria and explains joint supplies and demands, a concept which Marshall would take up later on. Further, he has deeply influenced our theories of profit and rent by interpreting the entrepreneurial gain as rents of differential ability. Indeed, Mangoldt definitely anticipates Schumpeter's theory of the entrepreneur. He clearly distinguishes profit as an independent category of income from interest (of the capitalist), by stressing different elements of gain such as the compensation for risk-bearing or for new goods or techniques of production and sale. The pioneer function of the entrepreneur, motivated also by intangibles, as in Smith's concept, is clearly expressed in the statement

> . . . die Auffindung und Verwirklichung der besten Produktionsmethoden . . . die Ausbeutung der von der Natur gegebenen Hilfsmittel, die Herstellung der Güter in der für das Bedürfnis dienlichsten Weise [the discovery and realization of the best methods of production, . . . the exploitation of natural resources, the manufacturing of goods in a way that is most appropriate for the need]. (1855, p. 68)

This means, of course, novel improvements as well.

Unfortunately, Mangoldt did not attempt to make these realistic and dynamic elements an essential part of his price theory via a notion of *evolutionary* competition as did Smith and Schumpeter. Thus he neglected the different properties and intensities of competition in different stages in the life cycle of a product.

Furthermore, his contribution to allocation theory was as pioneering as his analysis of coalitions on the labour market. Here he objected to profit participation by workers without risk-sharing. Finally, it is notable that Mangoldt originally extended Ricardo's theory of comparative costs by applying, although in rather vague terms, the notion of elasticity of demand and supply in the theory of international trade.

Mangoldt was, no doubt, one of the eminent theorists and rare pioneers in the 19th century whose achievements are still underrated and whose use of mathematics seems to be rather overrated. Though an abstract thinker, he seldom lost the binding ties to reality.

Selected Works

1847. *Über die Aufgabe, Stellung und Einrichtung der Sparkassen*. Dissertation, Tübingen University.

1855. *Die Lehre vom Unternehmergewinn: ein Beitrag zur Volkswirtschaftslehre.* Leipzig: Teuber.

1863. *Grundriss der Volkswirtschaftslehre,* 2nd ed. Stuttgart: Maier. A chapter was translated as 'The exchange ratio of goods'. *International Economic Papers* 11(1962): 32–5911.

Bibliography

Recktenwald, H.C. 1951. Zur Lehre von den Marktformen. *Weltwirtschaftliches Archiv* 67 (2): 298–326.

Recktenwald, H.C. 1985. Über das Selbstverständnis der ökonomischen Wissenschaft. In *Jahrbuch der Leibniz-Akademie der Wissenschaften und der Literatur*. Wiesbaden: Steiner.

Recktenwald, H.C., and P.A. Samuelson. 1986. *Über Thünens 'Der isolierte Staat'*. Darmstadt/Dusseldorf: Wirtschaft und Finanzen.

Manoilescu, Mihail (1891–?1950)

Nicholas Georgescu-Roegen

Mihail Manoïlescu was born in Tecuci (Romania) in 1891, the son of two elementary school teachers. His continuous school successes bespoke of the exceptional qualities on which he was to rely for political ambitions as well as for scholarly endeavours. As a top student at the Bucharest Polytechnical School, from which he obtained the engineer diploma in 1915, Manoïlescu became acquainted with the Crown Prince (the future Carol II) while the latter was attending the same class. It was a political asset subsequently enhanced when in 1930 Manoïlescu was instrumental in bringing Carol back from self-imposed exile. Manoïlescu started his reputation as a keen engineer and an astute operator by designing a better cannon and by organizing a successful Industrial Exhibition (1921). He began his catapulting political career by joining the People's Party and thereby becoming Under-Secretary of Finance (1926/27). Thereafter, he switched to the Peasant Party and, after Carol's return, he became in succession Minister of Communications, of Industry and Commerce, and Governor of the National Bank. In 1932, he occupied the newly established chair of political economy at the Polytechnic School.

Alongside that arduous activity he devoted time to his second attraction in life, the political and economic problems which he tackled with the characteristic subjective originality of the non-professional who, as Schumpeter remarked, often sees what the others do not. Several minor articles in Romanian as well as the few papers in foreign periodicals were thin and verbal. In his inaugural lecture (1932) and in his course (1938) he praised the mathematical tool and yet he criticized the Lausanne School for using the concept of *homo economicus* with measurable attributes. And in one corner, he intimated that great personalities decrease the entropy of the community. Worthy of note for his political evolution was his lecture on Neo-liberalism in a signal series of eighteen others on political doctrines (1923). His theory of industrial protectionism which attracted no little attention grew out of the economic situation of Romania, then a predominantly agrarian country facing increasingly adverse international price scissors. Manoïlescu laid bare the root of his philosophy in a 1928 lecture at the Société de Géographie when he stunned many of the French political elite by asserting that no nation, just as no individual, can get rich without exploiting the labour of others. (Almost certainly, he was citing neither John Locke nor Robert Owen.) And he topped it by the quip that the British always placed a gratis copy of the *Wealth of Nations* in each bale of cotton goods exported to India. The English translation of the 1929 French original, *The Theory of Protection* (1931), was reviewed in the foremost periodicals. Only Jacob Viner (1932), 'that unbelievably skillful advocate' of the classical doctrine (as B. Ohlin judged him), threw Ricardo's book at the author. But, as E. Hagen (1958) observed in his balanced examination, Viner did not deal with 'Manoïlescu's argument' which did not refer to the static model of an economy that just strives toward its best comparative advantage compatible with *given, immutable* possibilities. His argument was about the advantage not only of industrialization for an agrarian and overpopulated economy but also of industrialized nations able to purchase food for less labour than that of its exports (a point laid out in 1920 by J.M. Keynes in *The Economic Consequences of the Peace*). He thus went beyond the protectionism of infant industries favoured by J.S. Mill and Alfred Marshall; as he claimed (Preface, 1929), he did better than F. List. Only Ohlin recognized Manoïlescu's merit in his review (1931) and more pointedly in his 1967 monograph. Viner kept decrying Manoïlescu's thesis, as he did in discussing a 1946 independent argument of L.H. Bean (Viner 1952). Yet even Viner (1937) finally weakened, as he sought to justify the classical doctrine by the difference in occupational disutilities. The few theorists who mentioned Manoïlescu's argument lost themselves in a maze of unrealistic and inadequate assumptions. But by now hardly any economist

would deny that industrialization was responsible for economic development from the Tennessee Valley to Korea. Of course, the classical doctrine stands correct as concerns mineral deposits or special climates, but the point has been usually overlooked, as it was by Manoïlescu (1943) in his comparison of labour productivity between oil and other Romanian industries.

Between the two World Wars an atmosphere of dissatisfaction with capitalism reigned all over Europe. In 1936, none other than J.M. Keynes stressed 'the objectionable features of capitalism . . . the inequitable distribution of wealth and income'. As far back as 1848, J.S. Mill recognized its 'crass materialism', even the superiority of pure communism. The interwar circumstances kindled the interest of many economists (R. Kaula, J.M. Clark, G. O'Brien, and A.J. Penty, as examples) for the just price of the medieval principles and the guild system in which the individual had less freedom but was more secure and in which not only the producer but also the consumer was protected. All were only harking back to such authorities as Lujo Brentano, William Ashley and Max Weber. And as in Italian 'guild' is *corporazione* (in French, *corporation*), in the 1930s one Italian political scientist after another (e.g. Ugo Spirito, Filippo Carli, Carlo E. Fermi, and even the eminent mathematical economist Luigi Amoroso) identified fascism with a corporatist state. Manoïlescu (1934) saw in corporatism the imperative doctrine of this century just as, in his opinion, liberalism was for the 19th century. A long argument with numerous quotations from many authors and especially from Benito Mussolini did not produce a satisfactory definition of corporatism, nor of corporation. He opined that even Italian facism could not serve as an instance of a corporatism integral and pure, supposed to represent 'the functional organization of the nation . . . in its supreme interest' (pp. 80, 176). The quotation from Bernard Lavergne, 'The French people are actually represented in the Parliament, but France is not', characterizes the way Manoïlescu envisioned his subject. Curiously though, Manoïlescu presented a summary of his volume at the Stresa meeting of the Econometric Society in 1934.

As an almost natural sequel, there followed his *Le parti unique* (1936), which removed all doubts about the author's sympathy with the dictatorial regimes of the Hitler–Mussolini type. Manoïlescu was certainly not a member of any nazi party, but in this volume he spoke exactly as one, having only praise for all dictatorial chiefs. And he vaunted himself as the chief of 'The National-Corporatiste League' he had founded in Romania. It was a far cry from his 1923 'Neo-liberalism'. The volume has nonetheless some value for its information about dictatorial regimes from the USSR to Portugal and, more interestingly, about the incipient dictatorial parties all over Europe from Norway to Switzerland. It was because of the temper of the 1930s that this volume (just as the other two) was translated into several languages.

As the political tension worsened Carol, counting on Manoïlescu's being in favour with the Axis Powers, swore him in as Minister of Foreign Affairs in 1940. However, when Manoïlescu had to bow to the Vienna *Diktat* by which the Axis allotted a part of Transylvania to Hungary, the careers of both men were brought to a sudden end. Some time in 1948, the Communist regime threw Manoïlescu without trial in to jail where he died, presumably in 1950. His life typifies the fate of many intellectuals during the stormy political circumstances of that time in Europe.

Selected Works

1923. Neoliberalismul. In Institutul Social Român, *Doctrinele Partidelor Politice*. Bucharest: Cultura Naţională.

1929. *The theory of protection and international trade*. London: King, 1931. (Translation of *Théorie du protectionism et de l'échange international*. Paris: Giard.)

1932. La méthode dans l'économie politique (Inaugural lecture). *Bulletin de Mathématiques et de Physique Pures et Appliquées de l'Ecole Polytechnique Roi Carol II*, No. 4.

1934. *Le siècle du corporatisme: doctrine du corporatisme intégral et pure*. Paris: Felix Alcan. (Translation of *Secolul corporatismului*. Bucharest: Ciornei.)

1935. Arbeitsproductivität und Aussenhandel. *Weltwirtschaftliches Archiv*. (Rebuttal to J. Viner's review.)

1936. *Le parti unique*. Paris: Oeuvres Françaises.

1938. *Incercări in filosafia ştiinţelor economice*. Bucharest: Imprimeria Naţională.

1943. Productivitatea si rentabilitatea. *Enciclopedia României*, vol. 4. Bucharest: Imprimeria Nationalã.

References

Reviews of Manoïlescu. 1929. Include: A. Loria, *Revista Bancaria*, 1930; J. Viner, *Journal of Political Economy* 40(1): 121–125, 1932; L. Pasvolsky, *American Economic Review* 22(3): 477–478, 1932; J. Condliffe, *Economic Journal* 43(169) 143–145, 1933.

Corden, W. 1965. *Recent developments in the theory of international trade*. Princeton: Princeton University/ Department of Economics.

Hagen, E. 1958. An economic justification of protectionism. *Quarterly Journal of Economics* 72: 496–514.

Myint, H. 1963. Infant industry arguments for assistance to industries in the setting of dynamic trade theory. In *International trade theory in a developing world*, ed. R. Harrod and D. Hague. London: Macmillan.

Ohlin, B. 1931. Protection and non-competing groups. *Weltwirtschaftliches Archiv* 33: 30–45.

Ohlin, B. 1967. *Interregional and international trade*, revised ed. Cambridge, MA: Harvard University Press.

Schmitter, P.C. 1978. Reflections on Mihail Manoïlescu and the political consequences of delayed-dependent development in the periphery of Western Europe. In *Social change in Romania, 1860–1940*, ed. K. Jowitt. Berkeley: Institute of International Studies/University of California.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.

Viner, J. 1952. *International trade and economic development*. Glencoe: Free Press.

Mantoux, Paul (1877–1956)

R. Forster

Paul Mantoux published *La Révolution industrielle au XVIIIe siècle* in 1906. Since the first translation into English in 1927, the book has gone through not less than sixteen English impressions, the most recent in 1983. In his introduction to the 1961 edition, T.S. Ashton attributed the longevity of Mantoux' book to the author's clarity of thought and expression, his openness to amendment by new evidence, and his scrupulous objectivity. Mantoux' contribution to English economic history can be compared to that of his compatriot, Elie Halévy, to English political and social history. Both historians intended their synthetic accounts for a French audience, but they were even more valuable to English readers because of their freshness of approach and their liberation from national preoccupations and prejudices. Given the prodigious historical literature on the English Industrial Revolution, the continued serviceability of Mantoux' book is extraordinary. Even today, eighty years after the first French edition, Ashton's high praise remains pertinent: '. . . by far the best introduction to the subject in any language . . . a permanent work of reference' (Introduction to the 1961 edition, p. 27).

Of course research since 1906, and especially since 1927, has qualified or corrected many of Mantoux' interpretations. Mantoux' work predates the enormous influence of national-income economics on economic history associated with W.W. Rostow (ed.), *The Economics of 'Take-Off' into Sustained Growth* (1963), and the work of American growth economics especially since the 1950s (R. Fogel and D. North). In England a whole generation of historians, signalled by the three-volume work of J.H. Clapham in the 1930s, employed quantification systematically to measure all aspects of economic growth, including changes in per-capita real wages, and to resolve the 'standard of living' controversy once and for all (P. Deane and W.A. Cole, R.M. Hartwell, T.H. Ashton, A.J. Schwartz, W.O. Henderson, among others). But by the late 1960s, the Fabian Socialist tradition, now ably defended (and amended) by Christopher Hill, Eric Hobsbawm and Edward Thompson, reasserted itself, arguing that aggregate economic statistics only disguised the more fundamental issues of the quality of life in a working class culture.

In a curious way, the historical debates on the English Industrial Revolution have now moved closer to Mantoux' synthesis of 1906. This is not so much because of Mantoux' sympathy for the English Fabians when he was a young professor at the University of London before World War I, but rather because his mode of history-writing was essentially descriptive and developmental rather than hypothesis-testing or statistically oriented. On the other hand, although Mantoux emphasized that the Industrial Revolution 'gave birth to social classes whose progress and mutual opposition fill the history of our times' (1920s), the greater weight of his book was 'to show the continuity of the historical process underlying even the most rapid changes' (p. 476).

Recall that his was a history of the industrial revolution in the *18th century*, and that a good third of the book treats the 'evolution of traditional industry', English commercial growth from the 17th century, and the agricultural revolution of the 18th century – all 'preparatory changes' before the rash of inventions and the beginnings of the factory system at the end of the 18th century. Mantoux regarded the First Industrial Revolution as the culmination of trends in the English economy at least two centuries in the making. Mantoux had great respect for such venerable English institutions as the Poor Law and the humanitarian response of evangelical groups to child labour and other abuses of the new factory system. In the end, while he did not minimize the social dislocations created by the Industrial Revolution, Mantoux, not unlike David Landes in *The Unbound Prometheus* (1969), acclaimed this 'revolution' as an English achievement of momentous consequences – technological, economic, and social – for the entire world in the centuries to follow.

Paul Mantoux was not only an economic historian. He acted as the Interpreter of the Supreme War Council and of the Peace Conference at Versailles in 1919. A committed internationalist, he served as Director of the Political Section in the Secretariat of the League of Nations. Mantoux' article on 'Lost Opportunities of the League' in *World Crisis* (Geneva, 1938) struck a less optimistic tone than the one pervading his great work of 1906.

Selected Works

1903. *La crise du trade-unionisme*. Paris. English edn, New York: B. Franklin, 1971.
1906. *La révolution industrielle au XVIIIe siècle*. Trans. M. Vernon as *The industrial revolution in the eighteenth century*. London: Jonathan Cape, 1928. Numerous further editions.
1909. *A travers l'Angleterre contemporaine*. Paris: F. Alcan.
1919. *Paris conference, 1919: Proceedings of the council of four, march 24 – April 18, 1919*. Trans. J.B. Whitton. Geneva: Publications de l'Institut Universitaire des Hautes Etudes Internationales, 1964.

Manufacturing and De-industrialization

Ajit Singh

Industries in advanced countries have expanded and contracted in response to changes in technology and demand since the beginning of the industrial revolution. However, the phenomenon of de-industrialization, usually identified with the contraction of output or employment in the manufacturing sector as a whole, has only caused concern in these countries during the last decade or so. It has spawned a large academic literature, particularly in the UK, which was among the first industrial countries to manifest symptoms of de-industrialization (Singh 1977, 1982; Blackaby 1979; Beckerman 1979; Thirlwall 1982; Martin and Rowthorn 1986). It has also led to an important public debate on industrial policy in the UK, the US and in other industrial countries (for the UK, see Ball 1982; Eatwell 1982; Matthews and Sargent 1983; Singh 1979; Stout 1979; Godley and Cripps 1981; for the US, see Thurow 1980, 1984; Branson 1982; Krugman 1983; Schultze 1983; Norton 1986).

Significantly, 'de-industrialization' has long been a subject of great interest in developing countries like India. Several Indian economic historians have argued that as a consequence of free trade with Europe and economic policies of the colonial government, India suffered de-industrialization during the 19th century. Instead of industrial expansion, both the proportion of output and employment contributed by manufacturing industry fell in this period (see Bagchi 1976; for a sceptical view of this argument, see Little 1982; for a theoretical discussion, see Hicks 1969).

At an analytical level, the phenomenon of de-industrialization raises two central issues. Firstly, why should some industries rise and others fall, or more interestingly, why should the manufacturing sector as a whole decline absolutely, or in terms of its share of national output or employment? Secondly, does de-industrialization, defined in these terms, matter? As Singh (1977) asked in one of the first papers on de-industrialization: 'What is so special about industry that one should be concerned about de-industrialization? There has also been a considerable loss of employment from agriculture, but not much has been said, at least by economists, about de-ruralization.' In other words, the main question is whether de-industrialization can be regarded simply as a normal response to changing technology and tastes, or does it signify some structural disequilibrium in the economy as a whole with malignant overall consequences. The proponents of an industrial policy to correct de-industrialization in countries like the UK and the US clearly regard it as a manifestation of a structural disequilibrium.

Insofar as 'de-industrialization' in advanced countries is understood in terms of an absolute decline or a falling share of the manufacturing industry in total output and employment, the main stylized facts may be summarised as follows (Singh 1977, 1981, 1982; Blackaby 1979; Martin and Rowthorn 1986, which are the main sources of the figures below).

First, in the first decade or so of the long post-war boom (1945–73), the share of manufacturing in total employment increased in almost all industrial countries, including old industrial countries like the UK which already had a large proportion of their labour force employed in manufacturing. (As long ago as 1860, nearly 33% of the UK labour force was employed in manufacturing: Feinstein 1972.) The manufacturing employment share reached its peak (36.1% of civil employment) in the UK in 1955, but since then it has been in decline. The share fell marginally over the next decade (it was 34.8% in 1966), at a somewhat faster rate to 32.3% until 1973, and at a much faster rate since then (the 1981 figure was 26.4%). In countries like Belgium and Sweden, the pattern of change in the proportion of labour force employed in manufacturing was broadly similar to that of the UK, i.e. a slow decrease up to 1973 and an accelerated decline subsequently as the world economy slowed down after 1973. However, it is significant that in France and Germany, there was no evidence of a trend decline in the manufacturing industry's share of employment until 1973, although since then the share has fallen markedly in both countries. In Italy and Japan, the share actually expanded up to 1973, but even in these countries there has been a trend decline since then. In the US, the proportion of labour force employed in manufacturing fell from 28.5% in 1955 to 24.8% in 1973 and to 21.7% in 1981.

Second, in relation to the numbers employed in manufacturing and their rates of growth, manufacturing employment increased in all industrial countries in the post World War II period (including the UK) up to the mid-1960s. Between 1966 and 1973, although it fell in a few countries (e.g. the UK, the Netherlands, Belgium), manufacturing employment continued to grow in most OECD countries, including the US, Japan, Italy, France and Germany. However, during the decade 1973–83, manufacturing employment has fallen in almost all industrial countries, the rate of decline ranging from 0.04% per annum in Japan and the US to a massive 3.1% per annum in the UK and 3.4% per annum in Belgium. Thus, in the UK the numbers employed in manufacturing industry have fallen from their historical peak of 8.4 million workers in 1966 to 5.4 million at the end of 1984, a drop of 35%, half of which occurred in the five-year period 1979–84.

Third, turning to manufacturing production, the share of manufacturing in total production *at constant prices* did not show any trend decline in most industrial countries, including the UK, during the period of the long boom (1945–73). In the US, Japan, Italy, France and Germany, manufacturing's share at constant prices, reached its peak in 1973. However, between 1973 and 1983, the g.d.p. in all the leading industrial countries has expanded at an appreciably faster rate than manufacturing production, thus indicating a fall in the share of manufacturing in g.d.p. over this period. The abysmal record of the UK manufacturing industry again stands out: despite the slowdown in world economy after 1973, all the leading industrial countries other than the UK managed to expand their manufacturing production during the decade 1973–83, albeit at a far slower rate than before. However, manufacturing output in the UK in 1983 was 18% lower than it was a decade earlier.

Fourth, in contrast to the behaviour of manufacturing in constant prices, when the share of manufacturing in total production is measured *at current prices*, it shows a trend decline in a large number of industrial countries even before 1973. In the UK, the share fell from 36.1% in 1960 to 31.0% in 1973; in the US the corresponding decline was from 28.4 to 24.9%. The main reason for this difference in the constant and current prices shares of manufacturing lies in the fact that in many advanced countries because of the faster growth of productivity in manufacturing than in services, the terms of trade tended to move in favour of the latter.

In short, the empirical evidence indicates that before 1973, there was a small decline in the proportion of the labour force employed in manufacturing in some of the advanced countries. However, no country showed a trend decline in the share of manufacturing in output at constant price. Since the slowdown in world economic growth in 1973, the share of manufacturing in both output and employment has fallen to a greater or smaller degree in most industrial countries.

The falling share of manufacturing in total employment and output is predicted by the theorists of the post-industrial society or of the so-called 'service economy' as a natural outcome of the long run process of economic development (Fuchs 1968, 1981; Bell 1974; for a critical view, see Gerschuny 1978). Time series as well as cross-section studies of both developed and developing countries show that there is a striking similarity in the pattern of structural change as a society becomes more wealthy. Rising per capita incomes are accompanied by continuing falls in the share of the agricultural sector in national output and employment; the share of the manufacturing or of the broader category of industry as a whole, increases until a high level of per capita income is reached and then it begins to decline; the service sector's share continuously increases. Thus, on the basis of regression analysis of time series data for the US for over a century (1870–1978) as well as cross-section data for the OECD economies, Fuchs (1981) found that the share of industry in total output typically reaches its peak at a per capita income level of $3000 to $3500 (at 1972 prices). Both the OECD cross-section study and the US time series analysis yielded a broadly similar conclusion. (See also U.N. (1979), which is based on data from both advanced and developing countries.)

Analytically, the reasons for the observed changes in the sectoral shares of agriculture, manufacturing and services in the course of economic development lie in two factors: (a) the relative rates of growth of productivity, $\dot{p}$, and (b) the relative income elasticities of demand, η, for the products of these three sectors. η is usually much greater in manufacturing and services than in agriculture, where it typically has a value of less than 1; η is also thought to have a higher value in services than in manufacturing. On the other hand, $\dot{p}$ tends to be very much higher in manufacturing and agriculture than in services. (For instance, over the period 1929–65 in the US, Fuchs (1968) found that $\dot{p}$ in industry increased at the rate of 2.2% per annum, compared with 1.1% per annum in services. Such a large difference in $\dot{p}$ between industry and services over the long time period examined in this study, cannot be accounted for by the well-known problems of measurement of output in the service sector.) If these stylized facts about the relative values of η and $\dot{p}$ in the three broad sectors of the economy

hold, it is not difficult to see that as per capita g.d.p. increases, the share of services in output and employment will tend to increase first at the expense of agriculture and subsequently at the expense of manufacturing industry (Martin and Rowthorn 1986).

In considering the fall in the share of industry in the advanced countries in terms of the above analysis, there are two points which deserve attention. First, the observed decline in the share of manufacturing in total employment is much more due to the effects of differences in $\dot{p}$ than in η between manufacturing and services. In the US, for the period 1948–78, Fuchs (1981) ascribe three-fourths of increase in service sector employment relative to that in manufacturing to the much greater value of $\dot{p}$ in manufacturing than in services, and only one fourth to the relatively higher income elasticity of demand for services than for manufacturing. Secondly, it is important to bear in mind that although η may be higher in services than in manufacturing, as noted earlier, prices tend to rise more slowly in manufacturing than in services. This 'price effect' is likely to compensate in part for the disadvantage of manufacturing with respect to the 'income effect', η, so that overall there may not be much difference in the rates of growth of demand for the output of manufacturing and services. This suggests that in the long run in the advanced countries, manufacturing's share in output at constant prices is likely to fall relatively little compared with its share in employment. In the limit, the latter share (whose primary determinant, as noted above, is the sectoral differences in $\dot{p}$ may be expected to become as small as that of agriculture, i.e. tend towards zero.

If the fall in the industry's share of output and employment is simply due to structural features common to all modern economies, de-industrialization in this sense should not be regarded as a matter of serious concern. However, it is conceivable that for a particular economy, the extent of the fall in the share of industry in either output or employment may go beyond what may be expected at that economy's level of per capita g.d.p. To the extent that the manufacturing sector is regarded as an 'engine of economic growth' (Kaldor 1966; Cripps and Tarling 1973), such reduction in the size or contribution of manufacturing may be expected to lower the economy's future growth potential.

More importantly, Singh (1977) argued that in an *open* economy, the question whether de-industrialization can be regarded as a manifestation of structural disequilibrium in the economy, 'cannot be properly considered in terms of the characteristics of the domestic economy alone'. Such a proposition, he suggested, has a sensible meaning only in the context of the interactions of the economy with the rest of the world, i.e. in terms of its overall trading and payments position in the world economy. Since trade in manufactures is a major determinant of the current account balance of most advanced countries (usually much more important than the trade in services), an analysis of de-industrialization necessarily entails an examination of the performance of the country's manufacturing sector in the international economy.

With these considerations in mind, to give precision to the concept of structural disequilibrium, Singh defined an efficient manufacturing sector in the case of the UK economy, in the following terms: *Given the normal levels of the other components of the balance of payments*, an efficient manufacturing sector is one which (*currently as well as potentially*) not only *satisfies the demand of consumers at home at least cost*, but is *also able to sell enough of its products abroad to pay* for the *nation's import requirements*. This is, however, subject to the qualification that an 'efficient manufacturing sector' must be able to achieve these objectives *at socially acceptable levels*, of *output, employment, and the exchange rate*. The latter restrictions are extremely important since at low enough levels of output and employment, or more arguably at a sufficiently low exchange rate, almost any manufacturing sector may be able to meet this criterion of efficiency. (The exchange rate should be regarded here as an indicator of the acceptable levels of inflation and inequality of income distribution.) It is also necessary to emphasise the significance of the condition that, to be efficient, the manufacturing sector must be able to fulfil the

above requirements not merely currently, but also in the long run. For instance, a windfall gain to the balance of payments (e.g. from the discovery of North Sea oil) may put it temporarily into surplus (at desired levels of output, employment, etc.), although manufacturing industry may be incapable of ensuring this when 'normal' conditions return. Cairncross (1979) has rightly pointed out that an implication of this conception of an efficient manufacturing sector is that even when 'manufacturing output was actually growing in proportion to g.d.p. (as on one measure it did up to 1973), or even when manufacturing employment was growing in proportion to total employment', there may be de-industrialization, i.e. a structural disequilibrium in the sense of a progressive failure to achieve sufficient exports to pay for full employment level of imports at a 'reasonable' exchange rate.

There is a large body of evidence that during the last two decades the UK manufacturing industry has not only been 'inefficient' in these terms, i.e. characterised by long-term disequilibrium, but more importantly, this disequilibrium has been growing worse over time (Blackaby 1979; Singh 1977, 1982, 1986; CEPG 1976–82). It is important to distinguish the period before 1979 with the period after 1979. There are two important characteristics of the latter period which are significant: (a) a new Conservative administration came into office in May 1979 and embarked on a rather different set of economic policies than had been followed hitherto by successive governments of both the major political parties (Reddaway 1982); (b) from being a net importer of oil, Britain progressively became a significant exporter of oil.

Even before 1979, it was evident that, mainly because of the decline in the performance of the UK industry in the world economy, there had been a trend deterioration in the country's current account balance at full employment. Between 1964 and 1978, the UK's share of world manufacturing exports was nearly halved whilst the industry suffered a huge increase in import penetration in its own home market; significantly, this was despite the fact that over this period UK's costs and prices, expressed in terms of a common currency, had *fallen* relative to those in the competitor countries. The main long term problem of the economy before 1979 was that the current account was increasingly going into deficit well before full employment was reached. To illustrate, in 1965–66, the country was able to achieve a rough balance on the current account although there was near full employment (rate of unemployment of 1.5%). A decade later, in 1975, although nearly 4% of the labour force was unemployed, there was a current account deficit of £1700 million. Part of this was, indeed, due to the effects of the rise in oil prices in 1973. However, as CEPG (1976) showed, even assuming that the terms of trade had remained constant at the pre-1972 level, there would have been a current account deficit of £2000 million at full employment in that year. By 1977 the UK's current account deficit at full employment had soared to nearly £6000 million.

Disaggregated analyses show that the main reason for the above disequilibrium did not lie in the UK's trade in services or invisibles, but in visible trade in finished manufactures. Equally importantly, this deficit arose from the UK's trade with other advanced countries rather than with the Third World. With the latter, it in fact recorded a growing surplus (Singh 1982; see further section VI below).

Since 1979, the situation of the UK manufacturing industry in the world economy has deteriorated even further. The period of North Sea oil in the 1980s (by 1983, the UK's oil exports amounted to nearly 20% of her total merchandise exports) *could* have been used to restore the position of industry; instead, this period has coincided with industry's accelerated decline. In 1980, as a consequence of the North Sea oil, and the government policy of monetary and fiscal restraint with consequent high interest rates and economic slowdown, sterling appreciated sharply leading to an enormous squeeze on manufacturing industry. Manufacturing production fell in a single year by 9.3%, the largest such decline ever recorded in the UK in a twelve-month period – larger than those during the Great Depression after 1873 and after 1929. The maximum annual fall in manufacturing production in the first Great Depression occurred

between 1878 and 1879 and was 5.5%; that in the second Great Depression was 6.9% between 1930 and 1931 (Singh 1986). Manufacturing production fell further by over 6% in 1981. In 1982, for the first time in a century of its industrial history, Britain, the erstwhile workshop of the world, recorded a deficit on its manufacturing trade. In 1985, even after three years of economic recovery, the level of manufacturing output was $5\frac{1}{2}\%$ lower than in 1979 and 5% lower than in the first quarter of 1974 when, as a result of the miners' strike, the country was working only on a 'three-day week'. (For a fuller discussion of the relationship between oil, the government economic policy and de-industrialization, see Singh 1979; Forsyth and Kay 1980; Barker 1981; Coutts et al. 1986.)

In view of the fact that Britain's North Sea oil resources are limited, and their contribution to the balance of trade has most likely already reached its peak, the prospects for the future of the UK economy with a weak manufacturing industry are grim indeed. Coutts et al. (1986), on the basis of certain optimistic assumptions about the future expansion of world trade and competitiveness of British industry, project that Britain's current account balance would move from a surplus of £4 billion in 1985 (achieved at a rate of unemployment of 13%) to a deficit of £20 billion (at 1985 prices) in 1995 if non-oil output grows at a rate of $2\frac{1}{2}\%$ per annum (which is the minimum required for unemployment not to increase further). On the basis of the past econometric relationships, it is shown that much of this deficit will be due to an increase in the manufacturing deficit from £3 billion in 1985 to £23 billion (at 1985 prices) in 1995. Such a large current deficit, amounting to nearly 5% of g.d.p. in 1995, would clearly be unsustainable for any length of time, and most likely the government would have to lower the non-oil rate of economic growth to perhaps one and a half per cent per annum to achieve a satisfactory balance. This would imply a further sizeable rise in unemployment.

To sum up, the UK economy had been undergoing a long-term process of de-industrialization – in the sense of a progressive failure of the manufacturing industry to earn enough to pay for the full employment level of imports – even before the new Conservative government of Mrs Thatcher came into office in 1979. The economic policies of this government, despite the benefit of North Sea oil, instead of reversing this process, have managed greatly to accelerate it, with serious consequences for current and potential production and employment.

As for the reasons for de-industrialization in the UK, there are a number of passionately held views about the causes of this 'original sin', ranging from the laziness of British workers, the deficiencies of the educational system, the peculiarities of the English 'class system', the weaknesses of the managers, etc. However, there is no general agreement among economists on the reason or reasons for the poor performance of UK industry. In this connection, the Cambridge school of economists have put forward an important thesis (Singh 1977; CEPG, various issues; see also Cairncross 1979). It is argued that whatever the underlying cause of the long-term structure of disequilibrium of UK industry, if Britain continues to participate in the world economy on the same kinds of terms as before, and/or if it does not change the domestic production system, the long-term disequilibrium will keep on becoming more acute over time. This thesis is based on Myrdal's (1957) theory of circular and cumulative causation. It is suggested that the weaknesses of the British industry have led to a slow overall growth of the economy over the last quarter century relative to that of the competitor countries. This, in turn, is regarded as being responsible for the lack of dynamism in the country's productive system. Economies that grow quickly have higher investment, achieve faster technical progress, more product innovation and improvements in other important non-price spheres of competition. In addition, the take-home pay of workers in a fast-growing economy will generally also be growing more quickly. Other things being equal, this is likely to lead to better relations between workers and managers, with consequent benefits to productivity and performance. On account of its slow growth, UK industry has suffered on both these counts. The result has been a vicious circle of causation by which industry is increasingly unable to hold its own in either overseas or

home markets. The Cambridge economists further argue that because of the size of the structural disequilibrium of the UK industry, and the nature of the wage-price relationships in the economy, such disequilibrium cannot be corrected by currency depreciations.

Turning to the US, can the definition of an efficient manufacturing sector and of de-industrialization in the sense of structural disequilibrium of the economy (as outlined in section I) be applicable to that country as well? On the face of it, US industry is confronted with problems similar to those of UK industry: loss of world export markets, increasing import penetration and an enormous current account deficit brought about in large measure by the failure of industry to compete adequately either at home or abroad. A number of American commentators have argued that the US has been 'losing the economic race' (to use Thurow's (1984) phrase), particularly to Japan. The country is thought to suffer from institutional sclerosis which is reflected in declining supply elasticities and rising rates of core inflation (Scitovsky 1980). Kindleberger (1973) referred to 'the dynamic failure of the [US] economy to produce new exports to replace those now being eroded by the product cycle'. Similarly, Abramovitz (1981) asked in his presidential address to the American Economic Association in 1980: 'Can we mount a more energetic and successful response to the challenge of newly rising foreign competition after 1970 than Britain did after 1870?' More recently, the competitive failings of US industry have been documented in the *DRI Report on the US manufacturing industries* (Eckstein et al. 1984) and in the Report of the US President's Commission on Industrial Competitiveness (1985).

Nevertheless this view of the decline of US industry in the world economy is very much disputed in a number of other studies (Branson 1981; Lawrence 1984; Summers 1983; Bergstrom 1983; Norton 1986, which provides a very useful survey of this literature). Lawrence in particular argues that the problems of US industry in the 1980s stem largely from macroeconomic policies of the US government. He suggests that the fiscal deficit, leading to high interest rates and large capital inflows, has been responsible for the sharp appreciation on the dollar. This, together with the faster growth of US economy after 1982 compared with that of Europe, have led to the huge merchandise and current account deficits. Lawrence, however, estimates that a relatively small depreciation of the dollar, 'an improvement of less than 0.25% per year in relative US prices would suffice to ensure balanced trade in manufactured products'. The correctness of this optimistic view of US industry will be tested by events.

A widely held view about de-industrialization in the advanced countries of the North like the UK is that it is being caused by the industrialization of the South and particularly of the so-called newly industrialising countries (NICS), e.g. South Korea, Taiwan, Brazil, Mexico, India. It is suggested that the comparative advantage in a whole range of industrial products has shifted to the NICS and this has inevitably led to a reallocation of resources from the manufacturing industry in the North to other economic sectors as well as to the South through multinational investment (Beenstock 1984). In less sophisticated terms, the trade unions and business groups in the North ascribe the decline in manufacturing employment and the increase in overall unemployment in the industrial countries to the competition of cheap labour products from the South.

It is, indeed, true that during the 1960s and 1970s, the South achieved rapid industrial progress. The southern countries' share of world manufacturing production, although still quite small, increased by nearly 50% over the period 1960–1980, from 6.3 to 9.9%. Significantly, the South's share of world production rose not only in consumer goods but across a wide spectrum on industries, including capital goods products like steel, ship-building and engineering. Equally importantly, the South's share of world manufactured exports increased from 3.9% in 1960 to just over 5.0% in 1970 and to 9.0% in 1980. During the 1970s the advanced economies' imports of manufactures from the South expanded at twice the rate as their imports from each other (Singh 1982, 1984; UNIDO 1979).

It is important to emphasise that industrialization in one part of the world need not necessarily

take place at the expense of industrial development in another part of the world. Sayers (1965) made a useful distinction between 'complementary' and 'competitive' aspects of world economic expansion for any particular country or region. Economic growth elsewhere is 'complementary' to the extent that it raises demand for the country's exports, but it becomes 'competitive' insofar as it leads to the development of alternative sources of supply. So, from the point of view of a country or a region, the development of the world economy may be characterised by a changing balance between 'complementarity' and competitiveness'. During 1960–1980, not just the South, but the socialist countries of Eastern Europe industrialised very fast; the latter's share of world industrial production increased from 17% to 27% over these two decades. Nevertheless, it would be difficult to maintain that East European industrialization was achieved at the expense of the older industrial countries of Western Europe. On the contrary, most observers would agree that, if anything, West European industrial development was most probably positively helped by industrial demand from the East.

The question remains whether industrialization of the South, and in particular, of the NICS, has been more 'competitive' rather than 'complementary' to the North and, hence, responsible for the loss of manufacturing jobs in the advanced countries. A host of empirical studies answer this question in the negative (see OECD 1978, for a useful review of this research). More specifically, a number of these studies, particularly for the UK and the US, which follow a similar methodology, suggest, broadly, two kinds of conclusions. First, relative to the growth of productivity and changes in domestic demand in advanced countries, manufacturing trade as a whole (with both the developed and developing countries) has a relatively small effect on reducing manufacturing employment in these economies. Secondly, the net effect of manufacturing trade with the less developed countries on aggregate unemployment in the North, though negative, has been more or less negligible. Thus, the Foreign and Commonwealth Office (1979) concluded on the basis of an analysis of UK's trade in manufactured products with the NICS between 1970 and 1977 as follows: 'Any net displacement (of labour due to trade with the NICS) appears to have been quite small.' The main reason why the observed effects are so small is that although manufacturing imports from the NICS *ceteris paribus* reduces employment, the NICS have a very high propensity to import which leads to increased northern exports and, hence, employment. The overall effect tends to be slightly negative since the southern imports generally affect the relatively labour-intensive industries in the North whilst North's exports to the South occur in the relatively capital-intensive sectors.

Although these conclusions are not unreasonable, the underlying methodology of the above studies is open to serious reservations. The common analytical model on which they are based, makes changes in employment in an industry a function of changes in home demand, trade and productivity. Thus, in this research, increases in productivity *always* lead to a reduction in employment, which is clearly unsatisfactory. It is more acceptable to envisage that the growth of productivity leads to a reduction in prices which increases both domestic and export demand and, hence, generates a rise in output and employment. Moreover, the theoretical model does not take into account the fact that at least some of the advanced countries (e.g. the UK) during the reference period were balance of payments constrained. For such economies an increase in trade imbalance will have a multiplier effect on output and employment. The deterioration in the trade balance in a particular industry, or with a particular country, means that, *ceteris paribus*, unless there is an equal improvement of the balance in another industry, or with another country, the Government (through fiscal and monetary policies) is forced to run the economy at a lower level of output and employment than it otherwise would.

Singh (1981, 1982) used a rather different analytical model to study the effects of the UK's trade with the NICS over the period 1963–77. His methodology was more in keeping with the conception of de-industrialization discussed in section IV, i.e. the notion of long-term structural or trading disequilibrium. He concluded that

manufacturing trade was indeed the main cause of Britain's de-industrialization in this sense and, hence, responsible for the slow growth of output and employment in the economy. However, it was imbalance in the manufacturing trade with Japan and other advanced countries which was the chief source of this disequilibrium rather than Britain's trade with the Third World. Britain enjoyed trade surplus in manufactures with the NICS which had in fact progressively increased over the period studied, while with Japan and other advanced countries, Britain's trade balance had sharply deteriorated. Thus, despite the fast pace of industrialization in the NICS and the huge increase in their manufacturing exports to the UK, the UK's trade in finished manufactures (SITC 7 and 8) with the NICS during the 1960s and 1970s led to an increase in domestic output and employment rather than a decline. Nevertheless, Singh also found that there was evidence that in the case of some of the NICS, there were likely to emerge *in the future* imbalance in trade of a kind similar to UK's trade with Japan and other advanced countries, although this had not actually happened so far.

More generally, the OECD (1978) study for a group of advanced countries and a sub-set of NICS found that OECD's trade balance in manufacturing with NICS increased over the period 1970–77. The tentative general conclusion which emerges from this analysis is that even if in some balance of payments constrained advanced countries manufacturing trade with NICS had become disequilibrating in the 1970s, it is likely in general to have been a far smaller source of disequilibrium than the trade among the industrial countries themselves.

Turning to policies to reverse de-industrialization, there has been an intense economic policy debate on the subject on both sides of the Atlantic, particularly in the UK. With the UK's rates of unemployment during the mid-1980s approaching those recorded in the worst years of the 1930s Great Depression (if unemployment rates are measured on a comparable basis: see Tomalinson 1982), the central economic policy issue is how to reduce unemployment. It is generally agreed that a substantial sustainable increase in employment, other than through temporary make-work schemes, will require much faster economic growth than the trend rate experienced during the 1970s and 1980s. This in turn will necessitate a veritable reindustrialization of Britain – much faster growth of manufacturing production than has been achieved since 1973 – in order to sustain a current account balance at a time when the contribution of oil exports to the balance of payments will most likely be decreasing. (This is true despite the relatively greater contribution to the balance of payments which may be expected from the 'services' in the future; in view of the relative dimensions of manufactured exports and credits from services and the high income elasticity of demand for manufactures, the UK current account balance cannot be maintained at the desired growth rate simply by a faster expansion of services. See Singh 1977; Blackaby 1979; Coutts et al. 1986.)

Not surprisingly, however, there is no general agreement among economists about how faster economic growth can be realised, if at all. There are two major institutional parameters which constrain economic policy in Britain: (i) international economic arrangements which essentially consist of more or less free trade, free convertibility of currency and more or less free capital movements; and (ii) despite its set-backs in the 1980s, there is still a strong trade union movement with a major influence on wage-price determination. (In particular, the second constraint makes it difficult for a government to resort to the classic device of devaluation without risking inflation.)

There are a number of economists who believe that the long-term structural disequilibrium of the economy is now so acute that faster economic and industrial growth to reverse de-industrialization can only be achieved if one or the other of the above institutional constraints is relaxed. On the one side it is suggested that a further weakening of the power of the unions as well as more market-oriented supply-side policies are required to initiate a process of sustained economic growth. On the other, in a series of papers over the last decade, the Cambridge Economic Policy Group economists have argued that it is the international economic constraint which needs to be relaxed: Britain requires a long period of comprehensive

import controls as well as capital controls to permit sustained industrial and economic recovery without running into current account disequilibrium (Cripps and Godley 1978; CEPG 1976–82). These import controls are regarded as a necessary but not a sufficient condition for achieving faster economic growth. They would need to be supplemented with appropriate supply-side policies, which can, however, be either of the free-market or dirigistic variety depending on the political preferences of the policy makers (see Singh 1980, 1986, for a fuller analysis of these issues). The CEPG economists argue that the purpose of import controls is not to seek an improvement in the current account balance above its sustainable level but rather to expand production and employment by limiting the *propensity to import*. Over the medium term, it is shown that import controls in fact make possible a much larger volume of imports than under free trade. (The reason for this apparent anomaly is simply that the primary determinant of imports is the level of economic activity which would be much higher under a regime of import controls.) The CEPG economists suggest that this perspective should make it possible to negotiate a programme of control with Britain's main trading partners without inviting retaliation.

Several British economists, however, still believe that Britain can reverse de-industrialization within the present institutional framework of the economy (Hopkins et al. 1982; Matthews and Sargent 1983). They suggest a standard Keynesian policy of reflation coupled with currency depreciation and an incomes policy. However, the record of incomes policies in the UK during the last two decades has been unpromising (Tarling and Wilkinson 1977). Moreover, detailed empirical analysis (CEPG 1982) suggests that in view of the weaknesses of manufacturing industry and expected reduction in the contribution from oil exports, even if the incomes policy is reasonably successful, such a programme will run into balance of payments constraint at a relatively low rate of economic growth. These policies would not therefore lead to a significant reduction in unemployment; they would also be accompanied by a relatively high rate of inflation.

In the US, the economic policy debate concerning de-industrialization has centred around the so-called issue of 'industrial policy'. The advocates of an industrial policy (Thurow 1982; Bosworth 1983; Etzioni 1983) have suggested that the US needs major fiscal policy changes to raise the rate of investment and to modernise the social infrastructure as well as a targeted industrial policy to make the US industry regain its competitiveness. However, other US economists have questioned these arguments (Krugman 1983; Schultze 1983; Lawrence 1984). These scholars insist that the lack of competitiveness of the US industry in the 1980s has essentially been due to the appreciation of the dollar. Therefore, what is required in this view is more appropriate macroeconomic policies rather than an industrial policy. (For a recent comprehensive review of this US debate, see Norton 1986).

In conclusion, de-industrialization in the sense of falling levels or shares of manufacturing output and employment, need be of no greater concern than de-ruralization if it does not imply any structural disequilibrium of the economy which constrains it from achieving full utilization of resources or the desired growth of production. Agreeing with what he called the Cambridge view, Cairncross (1979) noted,

> a contraction of industrial employment is a matter for concern if it jeopardizes our eventual power to pay for the imports we need. . . . it is the loss of economic potential that is the crux of the matter. But whether that loss arises from the reasons given in Cambridge, whether it can be made good in the way propounded there, and whether it might yield to other, more familiar, but less agreeable treatment are matters on which there is not likely to be general agreement.

Bibliography

Abramovitz, M. 1981. Welfare quandaries and productivity concerns. *American Economic Review* 71(1): 1–17.

Bagchi, A.K. 1976. De-industrialisation in India in the nineteenth century: Some theoretical implications. *Journal of Development Studies* 12(2): 135–164.

Ball, R.J. 1982. *Money and employment*. London: Macmillan.

Barker, T.S. 1981. De-industrialisation, North Sea oil and an investment strategy for the U.K. In *Oil or*

industry, ed. T.S. Barker and V. Brailovsky. London: Academic.
Beckerman, W. (ed.). 1979. *Slow growth in Britain*. Oxford: Oxford University Press.
Beenstock, M. 1984. *The World economy in transition*. London/Boston: Allen & Unwin.
Bell, D. 1974. *The coming of post-industrial society*. London: Heinemann.
Blackaby, F. (ed.). 1979. *De-industrialisation*. London: Heinemann Educational Books.
Bosworth, B. 1983. Capital formation, technology, and economic policy. *Federal Reserve Bank of Kansas City Review* 231–259.
Branson, W.H. 1981. *Industrial policy and U.S. international trade*. In Wachter and Wachter (1981), 378–408.
Cairncross, A. 1979. What is de-industrialisation? In Blackaby (1979).
CEPG. 1976–82. *Economic Policy Review*. Department of Applied Economics, Cambridge.
Coutts, K. J., W. A. H. Godley, R. E. Rowthorn, and T. S. Ward. 1986. *A Cambridge bulletin on the Thatcher experiment*. Cambridge: Department of Applied Economics.
Cripps, F., and W. Godley. 1978. Control of imports as a means to full employment. *Cambridge Journal of Economics* 2(3): 327–334.
Cripps, T.F., and R.J. Tarling. 1973. *Growth in advanced capitalist economics*. Cambridge, MA: Cambridge University Press.
Eatwell, J. 1982. *Whatever happened to Britain?* London: Duckworth.
Eckstein, O., C. Caton, R. Brinner, and P. Duprey. 1984. *The DRI report on U.S. manufacturing industries*. New York: McGraw-Hill.
Etzioni, A. 1983. *An immodest agenda: Rebuilding America before the twenty-first century*. New York: New Press.
Feinstein, C.H. 1972. *National income, expenditure and output of the United Kingdom, 1855–1965*. Cambridge, MA: Cambridge University Press.
Foreign and Commonwealth Office. 1979. *The newly industrialising countries and the adjustment problem*. London: Foreign and Commonwealth Office, January.
Forsyth, P. J., and J. A. Kay. 1980. The economic implications of North Sea oil. *Fiscal Studies* 1(2).
Fuchs, V.R. 1968. *The service economy*. New York: National Bureau of Economic Research.
Fuchs, V.R. 1981. Economic growth and the rise of service employment. In *Towards an explanation of economic growth: Symposium 1980*, ed. H. Giersch, 221–242. Tübingen: J.C.B. Mohr.
Gerschuny, J. 1978. *After industrial society*. London: Macmillan.
Hicks, J.R. 1969. *A theory of economic history*. Oxford: Clarendon.
Hopkins, B., M. Miller, and W.B. Reddaway. 1982. An alternative economic strategy – A message of hope. *Cambridge Journal of Economics* 6(1): 85–103.
Kaldor, N. 1966. *Causes of the slow rate of economic growth of the United Kingdom*. Cambridge, UK: Cambridge University Press.
Kindleberger, C. P. 1973. An American economic climacteric? *The New York Times* 1.
Krugman, P. R. 1983. Targeted industrial policies: Theory and evidence. *Federal Reserve Bank of Kansas City Review* 123–155.
Lawrence, R.A. 1984. *Can America compete?* Washington, DC: Brookings Institution.
Little, I.M.D., et al. 1982. Indian industrialisation before 1945. In *The theory and experience of economic development: Essays in Honour of Sir W. Arthur Lewis*, ed. M. Gersovitz et al. London: George Allen & Unwin.
Martin, R., and R. Rowthorn (eds.). 1986. *The geography of de-industrialisation*. London: Macmillan.
Matthews, R.C.O., and J.R. Sargent (eds.). 1983. *Contemporary problems of economic policy*. London: Methuen.
Myrdal, G. 1957. *Economic theory and underdeveloped regions*. London: Duckworth.
Norton, R.D. 1986. Industrial policy and American renewal. *Journal of Economic Literature* 24(March): 1–40.
OECD. 1978. *Economic outlook, occasional studies*. Paris: Organization for Economic and Cultural Development.
Reddaway, W. B. 1982. The government's economic policy – An appraisal. *Three Banks Review* 136: 3–18.
Sayers, R.S. 1965. *The vicissitudes of an export economy: Britain since 1880*. Sydney: University of Sydney Press.
Schultze, C. L. 1983. Industrial policy: A dissent. *Brookings Review* 2(1): 3–12.
Scitovsky, T. 1980. Can capitalism survive? An old question in a new setting. *American Economic Review* 70(2): 1–9.
Singh, A. 1977. UK industry and the world economy: A case of de-industrialisation? *Cambridge Journal of Economics* 1(2): 113–116.
Singh, A. 1979. *North Sea oil and the reconstruction of the UK industry*. In Blackaby (1979).
Singh, A. 1980. Industrial policy and the economics of disequilibrium: A reply to Professors de Jong and Van der Zwan. In *Investeren en Worklooshеid*, ed. W. Hafkamp and G. Reuter. Brussels: Sampson Alphen a/d Rijn.
Singh, A. 1981. Third world industrialisation and the structure of the world economy. In *Microeconomic analysis; Essays in microeconomics and development*, ed. D. Currie, D. Peel, and W. Peters. London: Croom Helm.
Singh, A. 1982. *Structural changes in the UK economy: A long-term structural analysis of U.K.'s trade with less developed countries and its impact on the U.K. economy*. Vienna: UNIDO.
Singh, A. 1984. The interrupted industrial revolution of the third world: Prospects and policies for resumption. *Industry and Development* 12.

Singh, A. 1986. The long-term structural disequilibrium of the UK economy: employment, trade and import controls. In *Free trade – Managed trade? Perspectives on a realistic international trade order*, ed. G. Sjostedt and B. Sundelius. Boulder: Westview.
Stout, D.K. 1979. *De-industrialisation and industrial policy.* In Blackaby (1979).
Summers, L. 1983. Commentary. *Federal Reserve Bank of Kansas City Review* 79–83.
Tarling, R., and F. Wilkinson. 1977. The social contract: Post-war incomes policies and their inflationary impact. *Cambridge Journal of Economics* 1(4): 395–444.
Thirlwall, A. P. 1982. De-industrialisation in the U.K. *Lloyds Bank Review* 134: 22–37.
Thurow, L. 1980. *The zero-sum society.* New York: Basic Books.
Thurow, L. 1984. Losing the economic race. *New York Review of Books* 27: 29–31.
Tomalinson, J. 1982. Unemployment and policy in the 1930s and 1980s. *Three Banks Review* 135: 17–33.
UNIDO. 1979. *World industry since 1960: Progress and prospects*. New York: United Nations.
US President's Commission on Industrial Competitiveness. 1985. *Global competition: The new reality.* Washington, DC: GPO.
Wachter, M.L., and S.M. Wachter (eds.). 1981. *Towards a new U.S. industrial policy.* Philadelphia: University of Pennsylvania Press.

Mao Zedong [Mao Tse-Tung] (1893–1976)

Peter Nolan

Mao led the Chinese Communist Party (CCP) in its revolutionary struggle pre-1949 and was pre-eminent in the post-revolutionary leadership for most of the period from Liberation (1949) until his death in 1976. The degree to which Mao personally dominated China's post-revolutionary development is illustrated by the dramatic changes that have occurred since his death. It seems reasonable to speak of a 'Maoist model' to characterize China's development path for much of the period from 1949 to 1976.

There were a number of influences underlying this model. Nationalism was central to Mao's thinking. He was proud of China's historical achievements and angry at her humiliations in the century before 1949. He wished to build a powerful modern economy so that China would 'never again be an insulted nation'. China's cultural tradition permeated Mao's thought; his analysis of problems in terms of 'contradictions' owes as much to the traditional Chinese dialectic of *yin* and *yang* as to Marxism.

The Leninist–Stalinist application of Marxism in the USSR also influenced Mao (not always positively). From this tradition he accepted the notion of a post-revolutionary vanguard party overseeing all aspects of socio-economic life. From it too he absorbed the view of a 'socialist' economy as the antithesis of capitalism, i.e. no private ownership of the means of production and economic decisions determined not by market forces but by planners' administrative directions ('with us plans are primary and price is secondary . . . the law of value has no regulating function'). The adverse consequences of administrative planning under Mao were the same as those in economies with similar systems (e.g. low incentives for technical progress or to improve the range and quality of products; high incentives to hoard resources).

Mao was convinced of the possibility (and desirability) of changing popular consciousness, so that the main force motivating social action might become collective interests rather than personal gain. Although he wanted modernization and material progress, Mao stood outside the Marxist–Leninist tradition in thinking that 'socialist' values ('fighting self' and 'serving the people') might be more successfully developed among poor people ('poor people want change, want to do things, want revolution') and in the villages more easily than in the 'corrupting' cities. For Mao, Liberation marked the beginning of a long process of both economic development and 'permanent revolution' in China's class relations.

Mao's economic policies may be examined under four headings: (1) population; (2) economic growth; (3) rural institutions; (4) the international economy.

After 1949 Mao initially considered population control unnecessary. He was persuaded eventually of the problems of rapid population growth, but a sustained campaign to control population

growth was not implemented until the 1970s, so that China's population grew rapidly for most of the 1950s and 1960s.

Although Mao did not produce a rigorously formulated theory of economic growth, certain aspects of his thinking on this question can be identified. He considered a high rate of investment a necessary condition of rapid growth. Administrative planning via physical controls, direct control over the urban wage bill, and the CCP's influence on rural collectives' income distribution, together permitted a high rate of investment – China's 'accumulation' rate stood at over 30 percent of national income in most years from 1957 to 1976.

Mao's writings suggest that under him China broke away from the heavy industry emphasis of other 'socialist' countries. Unfortunately, the high investment rate, microeconomic inefficiency, slow technical progress and a vicious circle of self-expansion within the capital goods sector, together helped produce an alarming fall in the incremental output–capital ratio from the 1950s to the 1970s. From 1949 to 1957 heavy industry's investment share rose rapidly, and thereafter generally absorbed 45–55 per cent of state units' 'basic construction investment'. Many Chinese economists (when permitted) criticized the system that produced this result, but Mao refused to make the sweeping changes required to shift away from the heavy industry bias.

Mao considered microeconomic relationships to be important for economic growth. He argued that in a cooperative environment 'workers will look upon the enterprise as their own and not the cadres". This, he believed, would release the vast areas of human creativity left untapped by capitalism's antagonistic class relations. For Mao, a socialist enterprise was one in which workers had a powerful say in enterprise decision making, managers and technicians discarded their 'haughty airs' and participated in manual labour, the competitiveness of piece rates was replaced by time rates, differences in basic wages were kept within strict limits, and the proportion of income allocated 'according to need' rose over time. With the partial exception of Yugoslavia, these utopian ideas had not received such attention in the 'socialist' countries since the first months of War Communism in the USSR.

Despite their intrinsic problems, in a different setting such policies might have produced better results. However, in China they were often crudely applied (e.g. 'integrating' managers and technicians with ordinary workers by forcing them to wear dunces' hats in public) and were practised in enterprises with negligible independence and whose workers experienced little long-term growth of real income. These caused serious motivational problems.

The CCP led China's peasants through land reform and on to establish rural collectives which were the basic framework of economic activity for most Chinese people from the mid-1950s to the early 1980s. Mao thought they were an appropriate setting for developing 'socialist' values, avoiding the class conflict of 'capitalist' agriculture, and supporting disadvantaged peasants. He believed too that collectives would benefit from economies of scale. It was to prove much harder to develop 'socialist' values among peasants than Mao had anticipated. The CCP waged a constant, unsuccessful battle to 'cut off the tails of capitalism' in the villages. Moreover, in certain areas of farmwork (especially labour intensive crop cultivation) powerful managerial *dis*economies of scale appeared. Farm efficiency was adversely affected too by state control over key collective decisions, such as the allocation of income between accumulation and consumption. As a result, the micro-level problems were even worse in the countryside than in the cities.

Mao was afraid that extensive contact with the international economy would make China 'dependent' on outside forces. In the 1950s China built a comprehensive industrial system. Trade was viewed as a necessary evil. Exporting firms were denied direct contact with world markets; it made no difference to them whether their products succeeded or failed internationally. Unsurprisingly, China's export performance from the late 1950s to the late 1970s was poor. Mao did not wish China to have a high level of imports, confident that she could produce domestically most of the products she required and could be virtually self-sufficient in technical progress. He did not

permit foreign investment in China or China's acquisition of long-term debt. The economic costs of Mao's extreme position were high.

In the early 1970s, as China emerged from the isolation of the Cultural Revolution, Western economists were increasingly sympathetic to China. Development economics textbooks commonly included a brief section on the 'Maoist model'. While arguing against its transferability to different political systems, it was usually praised for its alleged achievements in combining quite rapid overall growth with the elimination of mass poverty and more equal income distribution than in most developing countries.

Since Mao's death, the mass of newly available statistical and anecdotal information has led to a major reappraisal of the Maoist epoch. There have been shocking allegations of mass starvation after the Great Leap Forward (1958–9) during which Maoist policies were applied in their purest form. China's official statistics show that its population fell by about 14 million from 1959 to 1961, suggesting a demographic disaster. Many Western observers enthused about Mao's utopian goals in the Cultural Revolution but it became clear that these had been pursued in a deeply repressive fashion, involving the imposition of one man's vision upon an increasingly unenthusiastic population. The end of Maoism was greeted with huge relief at all levels of Chinese society.

It can now be seen that the Chinese economy in the mid–1970s was in a state of crisis. Rapid population growth over two decades, an excessively high and unbalanced accumulation rate, pervasive microeconomic inefficiency, and isolation from the world economy, combined to produce little measured improvement in average living standards from the mid–1950s to the late 1970s, and in certain important respects (e.g. housing, cotton cloth, edible oil, entertainment) the situation had deteriorated. Despite some success in ensuring that a basic minimum consumption standard was provided, when Mao died there still were wide regional income disparities and a sizeable minority of the Chinese population was abjectly poor.

These problems were illuminated by the results of the post 1978 economic reforms, which dismantled many important aspects of the Maoist model. After 1978 average living standards rose dramatically and the proportion of the population in poverty declined sharply. It is impossible not to attribute these achievements (and important new problems) to the massive institutional reform (especially that in the countryside), the increased impact of market forces, expanded contact with the international economy, and alterations in the state's investment policy.

It is not surprising that the attractiveness of Mao's development model waned rapidly after his death. Perhaps the most fitting epitaph is that provided in 1978 by the elderly economist Chen Yun:

> Had Chairman Mao died in 1956, there would have been no doubt that he was a great leader of the Chinese people, a respected, loved and outstanding great man in the proletarian revolutionary movement of the world. Had he died in 1966, his meritorious achievements would have been somewhat tarnished but still very good. Since he actually died in 1976, there is nothing we can do about it.

Selected Works

Selected Works of Mao Tse-Tung, Vols. I–V. Peking: Foreign Languages Press.

Maoist Economics

Wei Li

Abstract

During the Maoist era (1949–76), China attempted to adopt Soviet-style central planning in her first Five-Year Plan, but soon changed course. Under the heavy hand of Chairman Mao Zedong, China created a unique style of central planning where the centre enunciated broad policy directives in the form of slogans that could be easily passed down to local cadres, who were given strong

incentives to find ways to implement them. This article outlines a consistent framework for analysing the policy changes during the Maoist era and their dramatic impact on the Chinese economy.

Keywords

Agricultural productivity; Agricultural taxation; Collectivization; Cultural Revolution (China); Famines; Great Leap Forward (China); Lysenkoism; Maoist economics; Marxist economics; Nutrition; Peasants; Planning; Tax compliance

JEL Classifications

P3

> Ten thousand years is too long; seize the day, seize the hour.
> ***(Mao Zedong, Mengjianghong – A Reply to Comrade Guo Moruo, 1963)***
>
> Had Mao died in 1956, there would be no doubt that he was a great leader of the Chinese people, a respected, loved and outstanding great man in the proletarian revolutionary movement of the world. Had he died in 1966, his meritorious achievements would have been somewhat tarnished but still very good. Since he actually died in 1976, there is nothing we can do about it.
>
> (Chen Yun at the Central Party Work Conference, November–December 1978. Quoted from Lardy and Lieberthal 1983. *Ming-Pao* (Hong Kong) 15 January 1979)

'Maoist economics' refers to the collection of economic policies implemented by the Communist Party of China (CPC) during the Maoist era, which began with the founding of the People's Republic in 1949 and ended shortly after the demise of Chairman Mao Zedong in 1976. Thanks to the CPC's meticulous cultivation of Mao's personality cult, Mao was able to exploit his 'mass line' political strategy by exhorting the masses to follow his vision when the CPC hierarchy was unwilling. As a result, Mao could set major policy initiatives with few checks and balances. But to attribute all major decisions to Mao would be an oversimplification, especially before 1958. The leadership of the CPC in Beijing and local cadres, often split into factions with different policy agendas and preferences (for a detailed historical account of the policy debates within the leadership circle in China in the 1950s and 1960s, see for example, Lardy and Lieberthal 1983; Riskin 1987; Bachman 1997), contributed not only to policy implementation but also to policy formulation. Maoist economics is therefore not synonymous with 'Mao Zedong Thought' on economic matters. ('Mao Zedong Thought' is considered an extension of Marxism–Leninism derived from the teachings of Mao Zedong and the distillation of the experience of the Communist revolution in China. It has been enshrined in the Constitution of the CPC as part of the party's official ideology since 1945. As China has embarked on market-oriented reforms since 1978, 'Deng Xiaoping Theory', which advocated the pragmatic concept of 'socialism with Chinese characteristics', has served as the party's working doctrine.)

The aim of this article is to outline a consistent framework for organizing and understanding the economic policies that were formulated and implemented during the Maoist era.

Agricultural Taxation and the Chinese-Style Central Planning

When on 1 October 1949 Chairman Mao proclaimed that the Chinese people had finally stood up at the ceremony for the founding of the People's Republic, China was a desperately poor agrarian economy ravaged by more than a century of internal turmoil, foreign invasions and civil wars. With most of her industrial assets either destroyed or looted by the Soviet forces that occupied Manchuria at the end of the Second World War, or removed to Taiwan and Hong Kong ahead of advancing Communist troops, China was 'poor and blank', as Mao (1958) put it. With 90 per cent of her population of 550 million living in abject poverty in the countryside and toiling on small plots of land using traditional labour-intensive farming technology, China was barely able to feed and clothe her population.

Since poor peasants made up the vast majority of the population, the CPC under Mao had focused on building its support base among

peasants by, among other things, promising to deliver what every peasant wanted: a private plot of land. Between 1946 and 1953, the CPC launched land reform, first in the territories under its control and then in all ethnically Chinese areas on the mainland after 1949. The process generally involved assigning each rural family a class status; motivating the poor, lower-middle, middle and initially the 'rich' peasants to engage in 'class struggle' against landlords; and expropriating land, draft animals, farm implements and property from landlords and redistributing them to landless peasants (Fairbank 1992). The 'class struggle', which included public trials, denunciations and mass executions of landlords and counter-revolutionaries, created an atmosphere of terror. But the land reform solidified support for the CPC among the poor and middle peasantry. (For an on-the-ground observation of the land reform in a Chinese village, see Hinton 1967.)

As the CPC secured military and political control of the mainland, its priority shifted to managing and rebuilding the war-torn economy. With the economy rebounding quickly, Chinese leaders turned their attention to long-run economic development, aimed at building a socialist, industrial nation. Having secured material and technical support from the Soviet Union, they adopted a Soviet-style, heavy-industry-oriented development strategy in the first Five Year Plan (FYP 1953–7). The plan called for massive industrial investment, including the construction of 156 industrial plants outfitted with imported Soviet equipment. The Soviet contributions to this big push included loans amounting to about four per cent of the total investment, technology transfers and 10,000 Soviet specialists (Fairbank 1992). The success of this ambitious plan therefore hinged on the ability of the government to mobilize investable surplus internally. Without a significant industrial and commercial sector, the government had to extract the needed surplus from the vast agricultural sector.

Throughout Chinese history, agricultural taxes, collected in kind, have been the primary source of government revenue. (Indeed the Chinese character for tax, *shui*, as a portmanteau of grain and convert, refers to a levy on the use of land payable in grain.) In the 1950s, China had a three-tiered agricultural tax structure. At the first tier was an in-kind levy on grain production, known as the 'government grain'. Peasants received no compensation for turning over the government grain to the State Grain Bureau. At a statutory rate of 15 per cent in 1950, this tax accounted for 39 per cent of government revenue. (This figure is calculated using data posted on the official website of China's Ministry of Finance.) In later years, as the price scissors – the differences between the prices on industrial and agricultural goods – widened, and as the industrial sector grew rapidly because of the massive capital expenditure funded largely by agricultural taxes, the share of explicit agricultural taxes dropped to six per cent by 1976.)

At the second tier was an implicit tax, a grain procurement quota, which dictated how much each peasant household had to sell to the State Grain Bureau out of their after-tax grain at below-market procurement prices. After meeting these two obligations, peasants would usually be left with just enough grain to sustain a subsistent living. Markets still existed in the early 1950s, where peasants could exchange some of their surplus produce for other goods. At the third tier was an inkind levy on rural labour. Under the traditional subsistence farming practice in China, peasants would take a break or work less intensively during agricultural offseasons in order to conserve food energy. To the government, this idling was unacceptable. Dams, irrigation systems, roads and other large-scale infrastructure projects could be worked on more intensively during off-seasons by drafting peasants to carry out backbreaking manual labour. Utilizing a mixture of exhortation and coercion, the government mobilized tens of millions of peasants for large construction projects in the 1950s.

Collecting the three-tiered taxes from hundreds of millions of independent peasant households was a daunting task. Tax enforcement became even harder when market prices of grain rose substantially in 1952 as a result of increased demand caused by rapid industrialization and urbanization and by the need to export agricultural products in exchange for Soviet equipment. In response, the government in 1953 closed the grain market and

monopolized grain trade by fiat, making it illegal for anyone other than the government to engage in large-scale grain trade. In 1954, it expanded the control to include oil seeds, cotton, pork, and other key agricultural commodities.

Extracting agricultural surplus was further hampered by the lower level of agricultural productivity in China than in more developed countries. With nearly 90 per cent of the population living in the countryside, China was producing barely enough food and wearable fibres to meet basic domestic needs. Estimates by Ashton et al. (1984) suggest that the daily average food energy intake in China in the 1950s was around 2000 calories per capita, below the 2350 calories recommended by the United Nations.

To further improve its extractive capability and to raise agricultural productivity, the government turned to collectivization. By organizing peasant households into collectives, the CPC could extend its political control down to the village level. The grass-roots party organizations could effectively monitor production to further improve tax compliance. Rooted in the prevailing ideology, Chinese leaders also believed that collectivization would enable peasants to take advantage of economies of scale, to learn best practices in scientific farming, to accelerate the adoption of high-yield seeds and modern inputs, and therefore to realize a great leap in agricultural productivity. (Apparently influenced by Soviet propaganda, Chinese leaders were taken in by the miraculous claims of productivity-boosting farming techniques made by a group of pseudo-scientists who dominated the Soviet agricultural science establishment. Because these pseudoscientific techniques contradicted the farming experience of Chinese peasants, the only way to propagate them was to make it a political task for rural collectives and grass-roots party organizations. For an account of Lysenkoism in the Soviet Union and its influence in China during the Maoist era, see Becker 1996.)

Collectivization, however, represented a radical reorganization of rural life in China. Given the importance of agriculture in the Chinese economy and the traumatic experience of forced collectivization in the Soviet Union (Becker 1996), China's first FYP emphasized voluntary participation and set out a relatively conservative and flexible timetable, calling for socialist transformation in agriculture to be accomplished in 10–15 years. Between 1952 and 1954, collectivization proceeded gradually. By 1954, only 11 per cent of peasant households were enrolled in elementary Agricultural Producers' Cooperatives (APCs), where members pooled their privately owned land, draft animals and large tools and used them jointly. APC members were paid wages for their labour as well as rents for their contributions in land and capital. While wage rates and rents were supposedly set at market levels, actual practice left many richer peasants complaining that the rents were insufficient. Reports of richer peasants exiting the cooperatives, selling and killing their draft animals, and downing trees on their plots in 1954 started to alarm leaders in Beijing. In January 1955, the CPC issued an urgent order for the protection of draft animals. The combination of state monopolization of grain trade and collectivization had, by the authorities' own admission, dampened the 'enthusiasm' of the peasants for production. Emergency measures that the government implemented included fixing procurement quotas and putting on hold any further push for collectivization in the spring of 1955. But any reprieve that peasants got was short-lived.

By the summer of 1955, imbalances in the economy from implementing the aggressive first FYP had reached record levels. The supply of agricultural products, raw materials and consumer goods could not keep up with the growing demand. With tax revenues insufficient to meet the funding needs in the first FYP, the government was running a large fiscal deficit. (In 1955, debt issuance by the government reached a record high of 2.5 per cent of GDP.) Factors that contributed to the imbalances included the agricultural bottleneck exacerbated by the collectivization movement, the ambitious first FYP that allocated massive investment to heavy industry, and the inherent difficulties of managing a centrally planned economy.

Mao's own analysis, however, identified overcentralization as a serious problem of the Soviet-style central planning whereby the planners tried to do what could be done better by local cadres. The solution that Mao put forth was not to stop the

expansion of the role of the state in the economy, but to limit the role of the nascent central planning bureaucracy and expand the role of local governments. He faulted the planners in Beijing for not doing enough to harness the enthusiasm of local cadres, workers and peasants for socialist transformation both in industry and in agriculture. In a policy speech delivered on 31 July 1955, Mao made the argument for accelerating socialist transformation in general and collectivization in particular.

> [Some] comrades fail to understand that socialist industrialization cannot be carried out in isolation from the cooperative transformation of agriculture. In the first place, as everyone knows, China's current level of production of commodity grain and raw materials for industry is low, whereas the state's need for them is growing year by year, and this presents a sharp contradiction. If we cannot basically solve the problem of agricultural cooperation within roughly three five-year plans, that is to say, if our agriculture cannot make a leap from small-scale farming with animal-drawn implements to large-scale mechanized farming, . . . then we shall fail to resolve the contradiction between the ever-increasing need for commodity grain and industrial raw materials and the present generally low output of staple crops, and we shall run into formidable difficulties in our socialist industrialization and be unable to complete it. (Mao 1977, pp. 196–7)

To ensure that Mao's vision was turned quickly into action, the CPC passed in October 1955 a resolution that reiterated the policy directive for accelerating collectivization and authorized the party hierarchy to criticize any party member who disagreed with the policy as a 'right-leaning opportunist'. ('The Resolution Regarding Agricultural Collectivization' was passed in the 6th Plenary Meeting of the 7th CPC Congress held in Beijing from 4–11 October 1955.)

As local cadres who moved decisively and quickly to implement this policy directive were publicly praised, and laggards were publicly criticized, local cadres found themselves locked into a rat race on who could coerce peasants to form bigger collectives at a faster pace. By the end of 1956, 96.3 per cent of all peasant households had joined collectives, more than 10 years ahead of the schedule set in the first FYP.

Mao's administrative decentralization was not a repudiation of the concept of central planning. It was an attempt to redefine central planning in the Chinese context with perhaps an implicit intent to enlarge the sphere of Mao's influence. By weakening the nascent central planning bureaucracy, Mao effectively strengthened his own influence in enunciating broad policy directives in the form of slogans that could be easily passed down to local cadres. To align the interests of local cadres with the centre, Mao offered high-powered incentives: those who found innovative ways to implement the centre's directives irrespective of economic consequences were rewarded with public praise and promotion, while those who ignored the centre's policy directives were punished with the humiliation of public criticism and denunciation. In more serious cases, those who resisted the centre's policy directives could be purged as 'rightists' or 'counter-revolutionaries'. Mao also made frequent use of brutal political campaigns against nonconformists and instilled an atmosphere of terror. (One of the most notorious political campaigns was the 1957 'anti-rightist' campaign; Fairbank 1992.) The resulting political system was one in which Mao could exploit his personality cult in enunciating broad policy directives without the inconveniences of checks and balances. Mao's administrative decentralization thus marked the beginning of the politicization of economic policy formulation and implementation in China. When Mao launched the Great Leap Forward (GLF) movement in 1958, the inaugural year of the second FYP, there was hardly any dissenting voice.

The Great Leap Forward

By setting production targets even more aggressively in the second FYP, the CPC hoped that China would grow out of the imbalances created during the first FYP by exhorting local cadres and the masses to make selfless sacrifices in order to transform China into an industrial, socialist nation. In March 1958, the CPC issued a new directive, calling local cadres to amalgamate smaller cooperatives into larger ones. Zealous local cadres in Henan province created township-sized collectives, dubbed 'People's

Communes'. Each of the communes was an all-encompassing institution that functioned as a local government, an agricultural collective, local government-owned industrial and commercial enterprises (one of the enduring legacies of Mao's administrative decentralization was the policy directive that encouraged the creation of local government-owned enterprises), local schools, and a militia integrated into the national defence system. In these collectives, communalization went beyond all means of production and invaded the private lives of peasants. For example, family kitchens were banned and were replaced by communal kitchens that offered members free meals (Li and Yang 2005). 'People's Communes are good because they are big and communal', declared Mao. By early autumn, communes had spread across China.

Believing that collectivization significantly boosted agricultural productivity, the CPC created a new rat race for local cadres by exhorting them to 'overcome reactionary conservatism' (*People's Daily*, 10 September 1958). Unable to deliver the expected increase in grain output, local cadres started to outdo each other in statistical gamesmanship by making wild claims about grain output. An initial tally of the 1958 grain output after the autumn harvest pegged it at 525 million metric tons (MMTs), up by nearly 170 per cent from 1957. The figure was subsequently revised down to a more modest 375 MMTs. (The downward revisions did not stop here. Two more were made: first to 250 on 22 August 1959 and then to 200 in 1979; Li and Yang 2005.)

With the numbers indicating that collectivization had permanently resolved China's agricultural bottleneck, the government raised agricultural taxes: grain procurement (including government grain) was increased from 46 million metric tons in 1957 to 64 in 1959; 16.4 million peasants, about twice the size of the industrial labour force in 1957, were relocated to cities in 1958 to support the expansion of industry and construction; and more than 100 million peasants were mobilized in the winter of 1957–8 to undertake large irrigation and land reclamation projects, and to operate millions of small 'backyard iron furnaces'. (Built using mud and bricks, these furnaces melted scrap metal – for example, iron woks made obsolete by communal kitchens – to produce iron, which even the government admitted was of useless quality; Becker 1996.) The increase in agricultural taxes allowed the government to raise national savings from 24.9 per cent of national income (measured by net material product) in 1957 to 43.8 per cent in 1959. These savings were almost exclusively invested in heavy industries (Riskin 1987, p. 142). Grain export was raised from an average of 2.11 million tons between 1953 and 1957 to 3.95 million tons in 1959 to meet payment obligations for importing capital goods.

The collectivization miracle was, however, a mirage. Lin (1990) finds that incentive problems within large collectives had deleterious effects on agricultural productivity. With actual grain output significantly lower than the falsified statistics, the agricultural taxes were excessive. Grain retained in rural areas fell sharply from 273 kg per capita in 1957 to 193 kg in 1959, and further down to 182 kg in 1960. Since grain was the primary source of food energy in China at the time, the drop in per capita food availability coincided with the onset of worst famine in human history. (Demographers who extrapolated mortality trends in China estimated the total number of premature deaths during the GLF famine at between 16.5 and 30 million; see Li and Yang 2005.)

As the disastrous consequences of the GLF policies became known in 1959, Mao temporarily stepped aside to let his pragmatic colleagues, Liu Shaoqi and Deng Xiaoping, take responsibility in managing both government and party affairs. The pragmatic leaders started to reverse course: they reduced grain procurement by ten million tons, increased agricultural labour force by more than 50 million between 1958 and 1962 by sending back new industrial recruits back to the countryside, dismantled communal kitchens, downsized the collectives and started to import grain. More importantly, they allowed spontaneous, bottom-up experimentation with market-oriented reforms in 1961. Grain output began to recover in 1961, but did not surpass its pre-GLF level of 195 million metric tons (recorded in 1957) until 1966, the first year of yet another political upheaval – the Cultural Revolution.

The Model

To better understand the trade-offs faced by Chinese policymakers and the key factors that contributed to the GLF disaster, I turn next to Mao's policy directive for accelerating collectivization with the aid of a simple two-sector dynamic model developed in Li and Yang (2005).

A key feature of the Li–Yang model is the explicit dependence of agricultural labour productivity on nutrition. For simplicity, assume that in the agricultural sector labour is the only factor and the technology exhibits constant returns to scale. If L_t is labour allocated to agriculture, the aggregate grain output in year t can be written as

$$Q_t = af(c_t)L_t \tag{1}$$

where $af(c_t)$ measures the contribution of nutrition to the labour productivity of an average worker who consumes c_t amount of grain in year t, and a is a productivity parameter. Experimental and empirical studies have found that $f(\cdot)$ tends to be an increasing, S-shaped function with $f''(c) \geq 0$ at a very low level of food intake, and $f''(c) < 0$ as food intake reaches a sufficiently high level. (For a survey on health, nutrition and economic development, see Strauss and Thomas 1998.) If the government taxes away p_t amount of grain output from each agricultural worker after the harvest in year t, the amount of grain saved for consumption in year $t + 1$ is then

$$c_{t+1} = af(c_t) - p_t \tag{2}$$

The industrial sector uses a Leontief technology that produces one unit of industrial output by employing one unit of labour, d units of capital service and m units of grain as an intermediate input in fixed proportions. Assume that all capital goods must be imported and paid for by exporting grain, and the exchange rate is one unit of grain to one unit of capital service. With abundant grain supply and the economy's labour supply normalized to 1, the industrial output is simply $1 - L_t$. The government is assumed to maximize a discounted flow of industrial output, $\sum_{t=0}^{\infty} \beta^t (1 - L_t)$, subject to the following budget constraint:

$$p_t L_t \geq (d + m + n)\ (1 - L_t), \tag{3}$$

where $\beta < 1$ is the government's discount factor and n is the food entitlement of each industrial worker. (For more discussion on food entitlement, see Li and Yang 2005. In 1956, the national average of monthly ration of grain for labourers assigned to the most physically demanding jobs was 25 kg. Retail prices of food items in stores were set by the government and played little role in resource allocation.) This constraint, which captures China's key bottleneck during the Maoist era, states that the amount of grain procured must be sufficient to meet export demand for the importation of capital goods, industrial demand for intermediate inputs, and food demand from industrial workers.

Given the government objective, the optimal solution calls for allocating just enough labour to grain production, so the constraint (3) is binding in each year. This implies that the optimal allocation of labour to grain production should be $L_t = (d + m + n)/(p_t + d + m + n)$. Substituting this binding constraint into eq. (2), one can show that the government's optimal policy is a solution to the following Euler equation for a given initial level of food consumption c_0:

$$a\beta f'(c_{t+1}) = \left(\frac{af(c_{t+1}) - c_{t+2} + d + m + n}{af(c_t) - c_{t+1} + d + m + n}\right)^2. \tag{4}$$

The optimal steady state policy is to set the food consumption per agricultural work $\bar{c}$ such that $f'(\bar{c}) = (a\beta)^{-1}$. The steady state is asymptotically stable if $f''(\bar{c}) < 0$. This stability condition is satisfied if the productivity effect of nutrition exhibits diminishing returns around the steady state per capita food consumption, which is consistent with previous experimental findings.

Under the stability condition, the steady state grain procurement $\bar{p}$ and industrial output $1 - \bar{L}$ are both increasing functions of agricultural productivity a and the discount factor β. The model therefore validates Mao's claim in his quoted policy speech that raising agricultural productivity would contribute to the relaxation of the agricultural bottleneck and hence permit a faster pace of industrialization. It also proves that patience is a

virtue: a more patient government, one that uses a larger discount factor β (or a lower discount rate) in setting intertemporal policies, can sustain a higher level of steady state agricultural and industrial production. The intuition is as follows. A more patient government, one that discounts future industrial production at a lower rate and is content with a lower growth rate, would set a lower tax rate on peasants, allowing them to improve nutrition and labour productivity. The improved productivity would in turn increase the tax base sufficiently high to more than compensate for any revenue loss from lowering the tax rate. As a result, both the grain procurement and industrial production are higher in the steady state for a more patient government.

Like Stalin, Mao was impatient. (In a speech delivered in 1931, Stalin 1952, used nationalistic rhetoric to demonstrate the imperative to press on with rapid industrialization regardless of the obstacles during the first 5 Year Plan of 1928–1932 in the USSR.) And, like Stalin, Mao saw collectivization as a means to achieve rapid industrialization. Expecting collectivization to raise a, the increasingly impatient planners exhorted local cadres to increase grain procurement and to divert more agricultural labour to industrial production and large infrastructure projects. Since collectivization actually caused a to fall, the GLF policies left many peasants with insufficient amount of grain for consumption. Malnutrition (and famine in several grain-producing provinces) significantly reduced labour productivity, leading to a collapse in grain production. The further reduction in grain output caused malnutrition, and famine to spread from the countryside into cities. The linkage between nutrition and productivity thus offers a dynamic explanation of why the negative incentive effect of collectivization could cascade into a major catastrophe. Empirical investigation by Li and Yang (2005) finds that the GLF policies were principally responsible for this disaster. As the GLF policies were reversed by Liu Shaoqi and Deng Xiaoping, the Chinese economy began to stabilize in 1962. In 1966, when the economy appeared to have fully recovered by 1966, Mao launched another political campaign – the Great Proletarian Cultural Revolution.

The Cultural Revolution

The post-GLF policies had some noteworthy features. First, the centre–province distribution of power was rebalanced in favour of the centre. The task of collecting reliable information on the prevalence and severity of famine was simply too important to be delegated to local cadres. Second, collectives were downsized by making village-level 'production brigades' responsible for their own finances, and communal kitchens were closed. More important, the policies permitted spontaneous experimentation with household responsibility schemes within collectives, allowed peasant families to keep small private plots, and reopened markets in which peasants could sell their surplus produce. These policies arrested the downward momentum and brought about a gradual recovery.

But they represented a humiliating retreat in the campaign towards socialism.

As long as the retreat was tactical, Mao was content standing on the sideline. However, Khrushchev's denunciation of Stalin's rule in 1956 and the subsequent de-Stalinization in the Soviet Union gave Mao reasons to be concerned about his own legacy. As soon as the economy recovered, Mao moved to reclaim power so that he could purge those who had the potential to become China's Khrushchev. In 1966, Mao turned against Liu and Deng. Exploiting his personality cult, Mao kicked off the 'Great Proletarian Cultural Revolution' in 1966 by exhorting the Red Guards, made up primarily of students and other urban youths, to rebel against the power base of Liu and Deng – the government and party hierarchy. Liu and Deng, along with many of their colleagues, were labelled 'capitalist roaders' and were purged in 1968.

Mindful of the fragile conditions in the Chinese countryside, moderate leaders did their best to keep the revolution from spreading into the countryside, preventing a rerun of the famine during the GLF. But the market-oriented reforms permitted under Liu and Deng were nullified. Agricultural productivity continued to stagnate until market-oriented reforms were restarted in 1978. The demand for food, however, continued to grow as a result of the post-war baby boom. Unable to raise grain procurement quotas to meet

the growing demand for food rations in the cities, the government resorted to sending millions of urban youths to the countryside to grow their own food and to receive 're-education'.

The Cultural Revolution brought politicization to every facet of life in China. It was better to be revolutionary (that is, loyal to Mao) than productive. Intellectuals and experts, considered less loyal to Mao, were sent to re-education camps in the countryside. Colleges were closed at first and were reopened later to admit only students from 'revolutionary families' – families of workers, peasants and soldiers – based on recommendations from grassroots party organizations. Seasoned bureaucrats and factory managers were purged by the Red Guards, and 'Revolutionary Committees', comprised of workers, peasants and students, took over government offices and state-owned enterprises.

The revolution paralysed the government and the nascent economic planning apparatus. With neither the plan nor the market to guide the allocation of resources, the economy fell into a state of anarchy. As coordination across regions fell by the wayside, regional self-sufficiency, a policy stance endorsed by Mao, became a necessity. Specialization based on regional comparative advantage gave way to the duplication of industrial structure across provinces. The economy stagnated until 1978, when a rehabilitated Deng Xiaoping restarted market-oriented reforms.

Discussion

One of the classic tenets of Marxian economics is that, with planning eliminating the 'anarchy of production', a planned economy can avoid or at least better manage large aggregate economic fluctuations (Ellman 1989). The experience of the Chinese-style central planning offers little support for this claim. The analysis of the GLF disaster by Li and Yang (2005) suggests that, on the contrary, central planning as practised in China exposed the economy to a new systemic risk. Because policy directives formulated at the centre had to be carried out in all localities, policy failures had generated large economic imbalances, severe economic and political crises, and prolonged stagnation. The source of the risk is the concentration of economic and political power in the hands of the planners. In the case of China, Chairman Mao, a charismatic leader, maintained a near monopoly on economic and political policies. With no effective checks and balances during the Maoist era, the economic and political system in China was incapable of arresting the momentum of apparently deleterious policy directives.

The Maoist era was tumultuous. It saw spectacular post-war reconstruction, the build-up of a rudimentary industrial economy aided by Soviet assistance, and the formation of a decentralized government administration that emphasized regional self-sufficiency on the one hand and economic collapse, stagnation, a personality cult and brutal 'class struggle' on the other. It conditioned a generation of pragmatic leaders who, after the demise of Mao, would restart market-oriented reforms through decentralized regional experimentation, disown the personality cult, ban mass movements and depoliticize economic policymaking, while resolutely maintaining the CPC's hold on power. The historical significance of Maoist economics may lie not in what it is but in what it is not.

See Also

- ▶ Agriculture and Economic Development
- ▶ China, Economics in
- ▶ Chinese Economic Reforms
- ▶ Command Economy
- ▶ Famines
- ▶ Planning

Bibliography

Ashton, B., K. Hill, A. Piazza, and R. Zeitz. 1984. Famine in China: 1958–61. *Population and Development Review* 10: 613–645.

Bachman, D.M. 1997. *Bureaucracy, economy, and leadership in China: The institutional origins of the great leap forward*. Cambridge, MA: Cambridge University Press.

Becker, J. 1996. *Hungry ghosts: Mao's secret famine*. New York: Henry Holt and Company.

Communist Party of China. *Constitution*. Online. Available at http://news.xinhuanet.com/ziliao/2004-11/24/content_2255749.htm. Accessed 13 June 2007.

Ellman, M. 1989. *Socialist planning*. Cambridge, MA: Cambridge University Press.
Fairbank, J.K. 1992. *China: A new history*. Cambridge, MA: Harvard University Press.
Hinton, W. 1967. *Fanshen: A documentary of revolution in a Chinese village*. New York: Monthly Review Press.
Lardy, N.R., and K. Lieberthal. 1983. Introduction. In *Chen yun's strategy for China's development: A non-maoist alternative*, The China Book Project, ed. N.R. Lardy and K. Lieberthal. Armonk: M.E. Sharpe.
Li, W., and D.T. Yang. 2005. The great leap forward: Anatomy of a central planning disaster. *Journal of Political Economy* 113: 840–877.
Lin, J.Y. 1990. Collectivization and China's agricultural crisis in 1959–61. *Journal of Political Economy* 98: 1228–1252.
Mao, Z. 1958. *Introducing a cooperative*. In Mao (1977).
Mao, Z. 1977. *Selected work of Mao Zedong*. Vol. 5. Beijing: Foreign Language Press.
Ministry of Finance, People's Republic of China. Online. Available at http://www.mof.gov.cn. Accessed 20 Feb 2007.
Riskin, C. 1987. *China's political economy*. Oxford: Oxford University Press.
Stalin, J.V. 1952. *Economic problems of socialism in the USSR*. Moscow: Foreign Language Publishing House.
Strauss, J., and D. Thomas. 1998. Health, nutrition, and economic development. *Journal of Economic Literature* 36: 866–817.

Marcet, Jane Haldimand (1769–1858)

Robert W. Dimand and Evelyn L. Forget

Keywords
British classical economics; Malthus, T. R.; Marcet, J. H.; Martineau, H.; Say, J.-B.; Utility

JEL Classifications
B31

The classical political economist Jane Haldimand Marcet was born in London, the eldest of ten children of Anthony (Antoine) Haldimand, a Swiss citizen who was a successful London banker and property developer, and his English wife, Jane Pickersgill. She was tutored at home, studying the same subjects as her brothers, and took charge of the household at the age of 15, when her mother died. In December 1799 she married Alexander Marcet, a London physician from Geneva. Since her father bequeathed all his children an equal share of the family fortune, regardless of gender, she was independently wealthy, with no need to write for money. Nonetheless, she wrote 30 educational books on chemistry, political economy, botany, mineralogy, grammar and history, many written in the form of conversations. Her first book, an introduction to experimental chemistry, was published in 1806 after attending Humphrey Davy's lectures at the Royal Institution and after repeating Davy's experiments at home in Alexander Marcet's laboratory. The book was adapted in the United States as a college text, and its tremendous commercial success is shown by the many plagiarized editions that emerged in a period with no effective international copyright law. It introduced the young Michael Faraday to science.

Jane Marcet encountered the ideas of Adam Smith through Sydney Smith's lectures on moral philosophy at the Royal Institution in 1804 and 1806. Alexander Marcet and David Ricardo were both elected to the Geological Society in 1808. Jane Marcet's younger brother, William Haldimand (who lived with the Marcets), was elected a director of the Bank of England in 1809 at the age of 25, and, like his sister, shared Ricardo's attribution of the rising price of bullion to the excessive issue of bank notes, which was very much a minority view among the directors of the Bank of England. James Mill and Thomas Robert Malthus were also friends of the Marcets. Jane Marcet's *Conversations on Political Economy*, published anonymously in 1816, attempted to make the economic ideas of Smith, Malthus, Ricardo and Jean-Baptiste Say accessible to a wider public. Robert Torrens declared her 'one female, at least, fully competent to instruct the members of the present cabinet in Political Economy', while J.R. McCulloch considered her book 'on the whole, the best introduction to the science that has yet appeared'. Ricardo's daughter read the book at her father's recommendation, and Say wrote for permission to 'translate sizeable passages from her excellent book' for his political economy class (Polkinghorn 1993, p. 55).

Jane Marcet (1816, 1833, 1851) was a successful popularizer of classical political economy, but she was also fully capable of independent judgement, sharing Ricardo's opposition to the Corn Laws rather than Malthus's support for them, and supporting the proposed Factory Act in 1833, contrary to the beliefs of her younger friend, Harriet Martineau. Marcet was more optimistic than Ricardo or Malthus about the prospects for economic growth, being less concerned that the working class would erode gains in the standard of living by heedlessly multiplying. Like Say, she placed more emphasis on utility than labour cost as a source of value: when Malthus, after high praise of her discussion of rent, protested that 'I think you have given too much sanction to Mr. Say's opinion reflecting utility', she cut out and discarded the rest of his letter (Polkinghorn 1986). A talented educational writer and the first woman to expound the principles of economics, Jane Marcet succeeded in bringing classical political economy (and other disciplines such as chemistry, botany and mineralogy) to a wider public.

See Also

- ▶ British Classical Economics
- ▶ Corn Laws, Free Trade and Protectionism
- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Martineau, Harriet (1802–1876)
- ▶ Ricardo, David (1772–1823)
- ▶ Say, Jean-Baptiste (1767–1832)

Selected Works

1816. *Conversations on political economy*, 3rd ed. London: Longman, 1818.
1833. *John Hopkins's notions on political economy*. London: Longman.
1851. *Rich and poor*. London: Longman.

Bibliography

Bodkin, R.G. 1999. The issue of female agency in classical economic thought: Jane Marcet, Harriet Martineau, and the men. *Gender Issues* 17: 62–73.
Polkinghorn, B. 1986. An unpublished letter from Malthus to Jane Marcet. *American Economic Review* 76: 845–847.
Polkinghorn, B. 1993. *Jane Marcet: An uncommon woman*. Aldermaston: Forestwood Publications.
Shackleton, J.R. 1990. Jane Marcet and Harriet Martineau: Pioneers of economic education. *History of Education* 19: 283–297.

Marchal, Jean (born 1905)

J. Lecaillon

Marchal was born on 25 June 1905 at Colombey-les-Belles, France. He received his doctorate in economic science with a thesis entitled *Union Douaniére et organisation européene*. He was professor at the University of Nancy in 1935, then at the University of Paris from 1948 to 1972. He was elected member of the Academy of Moral and Political Sciences in 1980.

His initial work on the problem of value led him to make provision for psychological, sociological and structural factors. This concern for realism is found in his studies of the price mechanism (1946) with the introduction of correspondences, of imperfect competition phenomena and diversity of periods of analysis (very short, short, long ...). The same concern reappeared in research on the distribution of national income (1954). This distribution is not considered to be the same as allocation of parts of the National Product to abstract factors of production but as the result of an impact between the social classes defined by their position in the economy.

Marchal has also studied the French monetary system (1964) and the international monetary system (1975).

Selected Works

1929. *Union douanière et organisation européene*. Paris: Sirey.
1946. *Le mécanisme des prix*, 4th ed. Paris: Librairie de Médicis, 1966.

1951. The construction of a new theory of profit. *American Economic Review* 41: 549–565. 1957. Wage theory and social groups. In *The theory of wage determination*, ed. J.T. Dunlop. London: Macmillan.
1958–70. *La répartition du revenu national*, 4 vols. Paris: Librairies techniques.
1964. *Monnaie et credit*, 3 vols, 8th ed. Paris: cujas, 1985.
1978. The spreading progress of incomes in an economy: A reassessment of the multiplier theory through the probabilistic approach. In *Pioneering economics: International essays in honour of G. Demaria*, ed. T. Bagiotti and G. Franco. Padua: Cedam.

Marczewski, Jean (Born 1908)

J. Lecaillon

Marczewski was born on 27 May 1908 in Warsaw. He enlisted voluntarily in France during the World War II, and was deported to Germany. On his return to France he became Professor at Caen University in 1950, then at the University of Paris from 1959 to 1977.

After working on monetary policy, where he shows the limits of 'deficit spending' (1941), he developed a critical analysis of central planning (1956), arguing that despite the progress of mathematical programming and information theory, planning does not seem able to replace the market. The application of national accounting techniques (1965) to historical research led Marczewski to underline the prime importance of 'historic variables', unique in time and space, which play the role of exogenous variables in the determination of economic relationships.

Research on inflation and unemployment (1978), tested on French and German examples, led him to analysis of stagflation. A generalization of the analysis of real and monetary flows which compose the circuits of international exchange led to a theory of the movements of expansion and recession of world economy (1984).

Selected Works

1941. *Politique monétaire et financière du IIIème Reich*. Paris: Sirey.
1956. *Planification et croissance des démocraties populaires*. Paris: Presses Universitaires de France.
1965. *Introduction à l'histoire quantitative*. Geneva: Librairie Droz.
1974. *Crisis in socialist planning*. New York: Praeger.
1978a. *Inflation and unemployment in France: A quantitative analysis*. New York: Praeger.
1978b. *Vaincre l'inflation et le chômage*. Paris: Economica.

Marget, Arthur William (1899–1962)

Donald A. Walker

Keywords
Cash-balance approach; Keynesian revolution; Marget, A. W.; Quantity theory of money; Velocity of circulation

JEL Classifications
B31

A leading monetary theorist during the first half of his career, Marget went on to make an even greater contribution by formulating and implementing government policies regarding international banking and finance. Born in Chelsea, Massachusetts, on 17 October 1899, he graduated from Harvard with AB (1920) and MA (1921) degrees in Semitics, and a Ph.D. in economics (1927). He taught economics at Harvard (1920–1927) and the University of Minnesota (1927–1943; resigned 1948). He died in Guatemala City on 5 September 1962.

As an academician, Marget's principal concern was with the central problems of monetary theory, and since these had been so strikingly shaped by John Maynard Keynes, much of Marget's work

became a critique of his views. Marget regarded himself as building upon an enduring neoclassical tradition, and saw the Keynesian Revolution as a largely misdirected episode that had the merit, however, of making some genuine contributions, and especially of stimulating the sort of re-examinations and refinements of doctrine exemplified by his own writings. His most significant critical contributions were his evaluations of Keynes's *Treatise on Money,* of liquidity preference, of Keynes's treatment of expectations, and of the implications of Keynes's *General Theory of Employment, Interest and Money* for the theory of prices (Marget 1938, 1942). Marget's principal positive contributions were an extension and refinement of the concepts of the velocity of circulation of money and of goods; his reformulation of the quantity equation relating prices, output and money; his argument that the cash-balance approach is useful only in connection with the analysis of changes in the velocity of the circulation of money; and his analysis of the relevance of particular demand curves and their elasticity to the structure of money prices (Marget 1938, 1942). The valuable elements of his critique of 20th-century theory and of his constructive writings have been assimilated into the discipline and are no longer a focus of discussion. Marget also undertook some studies in the history of thought which are among the best work on the subjects with which he dealt. Particularly worthy of note are his examinations of the monetary theory of 19th-century neoclassical economists (Marget 1931, 1935, 1938, 1942).

As an applied economist, Marget was concerned with international financial policies. While a major (1943–1945) and a lieutenant colonel (1945) in the US Army, he devoted himself to preparations to bring about the economic and financial rehabilitation of Austria. He then became chief of the finance division of the US Allied Command for Austria (1946–1949); a member of the US delegation in London that prepared for the treaty with Austria (1947); and a member of the US delegation to the Council of Foreign Ministers, which was charged with negotiating that treaty in London and Moscow (1947). His subsequent career included the positions of Chief of the Finance Division at the headquarters of the Marshall Plan in Paris (July 1948–December 1949), consultant to the US Treasury (1948), and Director of the Division of International Finance of the Federal Reserve Board of Governors (January 1950–April 1961). Among other activities, he represented the Board at meetings of the central banks of the western hemisphere in Bogotá (1956) and Guatemala City (1958). He then became the US representative to the Central American Common Market in Guatemala City and an adviser to the Common Market Bank in Honduras (April 1961–September 1962). In the latter roles he was instrumental in promoting the effectiveness of the Common Market policies.

Marget's scholarly work was distinguished by an insistence on logical clarity and an amassing of scholarly detail in the presentation of his expositions. His bureaucratic work was distinguished by an outstanding ability to suggest workable new financial institutions and procedures.

Selected Works

1931. Léon Walras and the 'cash-balance approach' to the problem of the value of money. *Journal of Political Economy* 39:569–600.

1932. The relation between the velocity of circulation of money and the velocity of circulation of goods. Parts I and II. *Journal of Political Economy* 40:289–313, 477–512.

1935. The monetary aspects of the Walrasian system. *Journal of Political Economy* 43:145–86.

1938, 1942. *The theory of prices: A re-examination of the central problems of monetary theory*, 2 vols. New York: Prentice-Hall.

Marginal Abatement Costs

Ross McKitrick

Abstract

The marginal abatement cost curve (MAC) shows, for every emissions level, a firm or

industry's marginal cost (foregone profits) of reducing emissions, or equivalently, its marginal willingness to pay for the right to emit one additional unit of pollution. The term 'abatement' here denotes a unit reduction in emissions, as distinct from units of specific abating inputs, such as smokestack scrubbers. The MAC is not invariant to the form of the emissions control policy. The aggregate MAC for an industry is only well defined in cases where the policy achieves cost efficiency by equalising MAC levels across all emitters.

Keywords

Abatement; Efficiency; Emissions; Environmental policy

JEL Classifications

Q50; Q52

The marginal abatement cost curve (MAC) shows, for every emissions level, a firm or industry's marginal cost (measured as foregone profits) from reducing emissions by one unit, or equivalently, its marginal willingness to pay for the right to emit one more unit of pollution. It represents, in effect, the firm's demand curve for emissions. The term 'abatement' as used here denotes a unit reduction in emissions, rather than a unit of specific abating inputs, such as smokestack scrubbers.

The MAC is not the same as the marginal cost of any one particular type of abatement equipment. Instead, it is derived as the profit-maximising response to the relaxation of a constraint on emissions, taking into account the availability of emission control devices as well as changes in output and input levels.

The MAC plays a key role in the definition of the optimal level of pollution, which occurs where the MAC crosses the marginal damages (MD) line. However, the MAC is not invariant to the form of the emissions control policy, since some policies are more costly than others at achieving the same reduction in emissions. Where regulators depart from economic instruments (emission taxes and tradable permits) in favour of command-and-control regulations, intensity targets, technology standards or other mechanisms that do not minimise compliance costs, the MAC rotates higher than it otherwise would.

For an individual firm, the MAC can be derived by introducing an emissions function into the description of the firm (see Derivation section below). Abating inputs are assumed to be non-productive (of output) and are used only to reduce emissions. If an abating input is also productive of output, it is assumed to be used already to its profit-maximising level, so further use for emissions-abating purposes would not be marginally profitable; hence there is no loss in generality by simply defining abating inputs as non-productive. In the dual optimisation approach, the firm chooses its optimal level of output and abating inputs subject to a constraint on total emissions. This yields an expression for the marginal profits from relaxing the emissions constraint, or in other words the marginal private value of polluting, which corresponds to the firm's demand curve for emissions. If all firms face the same marginal cost (price) for emissions, the industry MAC is the horizontal sum of firm MACs. If the policy does not yield equivalent MAC levels at the assigned emission reduction targets, then the horizontal summation is invalid and the industry MAC is not well-defined, as will be discussed below.

Unregulated emissions are located at the point where the MAC is zero. In response to an emissions tax the MAC shows the level of emissions chosen in response to each level of the tax. In response to a tradable permits policy it corresponds to the demand curve for permits. The area under the MAC over an interval shows the total cost to the firm, in reduced profits, of reducing emissions by the amount of the interval (or, alternatively, the total benefit of increased emissions). When the quantity of emissions is the constraint variable, the quantity times the associated MAC equals the scarcity rents created by the policy. When price is the constraint variable, the MAC times the associated quantity is the total tax revenue.

This article will present a derivation of the MAC, some key points regarding its interpretation and some empirical estimates.

Derivation

This section presents the derivation in McKitrick (1999), to which readers are directed for more detailed discussion (see also McKitrick 2010, Chap. 3). An example of a MAC is shown in Fig. 1. It is directly analogous to an ordinary factor demand curve, treating the right to emit a unit of pollution as the factor. Suppose a firm produces output y which sells at price p. The firm can employ a non-productive pollution control input a that costs q per unit. Emissions e are a function of output y and abating inputs a:

$$e = e(y, a). \tag{1}$$

For a fixed level of a, emissions are assumed to rise in y at an increasing rate. This arises from diminishing marginal productivity of inputs, which implies increased waste and pollution residuals. For a fixed output level y, emissions fall in a at a decreasing rate, in keeping with the assumption of diminishing marginal productivity of abating inputs; in other words, that the most efficient and cheapest options will typically be used first.

In Fig. 1, unregulated emissions are denoted $\bar{e}$ and occur at the point where the MAC is zero. We do not typically draw the MAC going below zero since that would imply that firms can increase profits by reducing emissions, and we assume they will already have done so if it is possible. If emissions are unregulated the firm sets $a = 0$, which means the profit-maximisation problem reduces to the standard one with a solution where price equals marginal cost. If we denote that output level as y^*, then by Eq. (1), $e(y^*, 0) = \bar{e}$.

The firm's cost function c is written

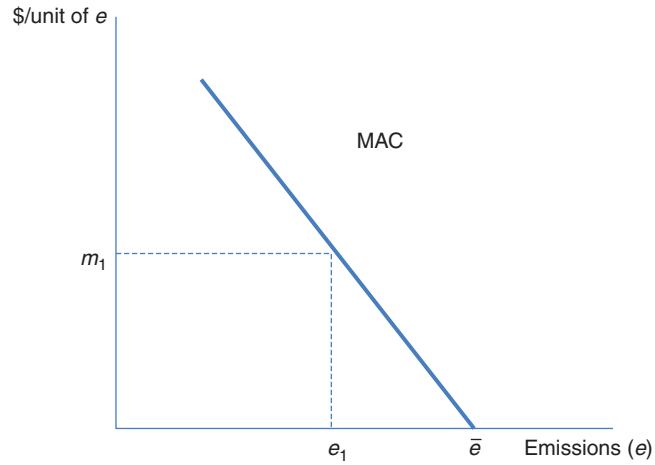
$$c = c(w, y, a) \tag{2}$$

where w is the vector of input costs including q. a appears in Eq. (2) because it is not a productive input, in other words it does not appear in the production function, so its corresponding price is not an element in w. In the absence of an emissions constraint the firm would set a to zero. If an abating input is assumed also to be productive, it must therefore be used already to its profit-maximising level, so without loss of generality we can simply define a as the amount of such an input used beyond its optimal level as a productive input.

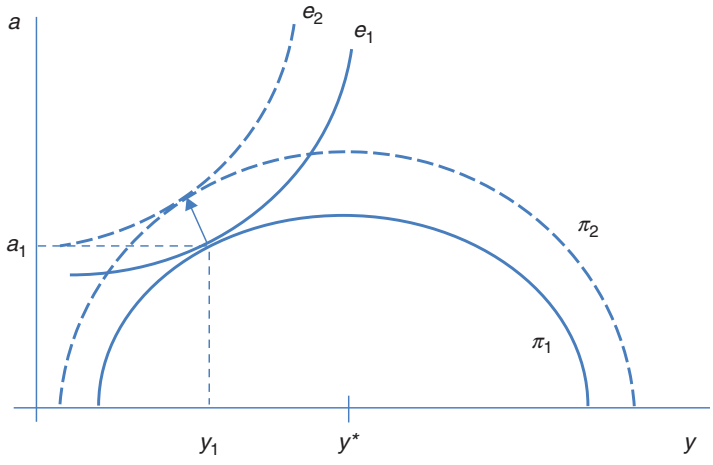
Profits are $\pi = py - c(w, y, a)$. The firm maximises profits subject to an emissions constraint

$$e(y, a) \leq e_1. \tag{3}$$

Figure 2 illustrates the resulting tangency in (y, a) space. The horizontal axis shows output (y) and the vertical axis shows pollution abatement equipment (a). The line labelled e_1 is an iso-emission



Marginal Abatement Costs, Fig. 1 Marginal abatement cost curve



Marginal Abatement Costs, Fig. 2 *Solid lines*: tangency between iso-emissions line for fixed emissions level e_1 and iso-profit line for fixed profit level π_1, showing optimal abatement and output levels denoted (a_1, y_1) respectively. *Dashed lines*: same for π_2 and e_2

line, showing combinations of y and a that yield a constant emissions level. The line labeled π_1 is an iso-profit line, showing combinations of y and a that yield a constant profit level. Profits go up by moving off an iso-profit line towards the horizontal axis at the point y^*, which is the optimal output level in the absence of emission regulations. So the firm seeks the lowest possible iso-profit line that still touches the emissions constraint line, which occurs at the tangency point (y_1, a_1).

Since emissions increase by moving off the iso-emissions line in an upward direction, a tightening of the emission standard is represented as a shift in the line e_1 back to e_2. The firm now moves to a lower profit level associated with the iso-profit line π_2, the tangency point for which implies more use of abatement equipment and less output. Hence the firm adjusts both output and abatement equipment levels in response to a changed emission constraint.

The MAC relates the change in profits ($\pi_1 - \pi_2$) to the change in emissions ($e_1 - e_2$). We can rearrange (1) into the form $a = a(y, e)$, showing the amount of the abating input needed at output level y to achieve emissions level e, and substitute this into the profit function to yield

$$\pi(p, y, e) = py - c(w, y, a(y, e)). \tag{4}$$

Assuming p is constant, the derivative of Eq. (4) with respect to e is

$$\frac{d\pi}{de} = -\frac{\partial c}{\partial a}\frac{\partial a(y, e)}{\partial e}. \tag{5}$$

This is the marginal abatement cost curve. In Fig. 2, a reduction in the allowed emissions level causes the firm to reduce output and use more abatement equipment. The cost to the firm is the change in profits $d\pi \equiv \pi_1 - \pi_2$.

When reading Fig. 1 from right to left, for instance from $\bar{e}$ to e_1, the area under the MAC denotes the total abatement costs (TAC) associated with the reduction in emissions. In Fig. 2 it would correspond to the difference between the unregulated profits associated with output y^* and the profit level π_1. The height of the MAC in Fig. 1 corresponds to the marginal cost of one more unit of emission reductions. Reading the same graph from left to right, the height of the graph corresponds to the marginal profits associated with one more unit of allowed emissions (Eq. (5)), and the area under the curve over an interval shows the total benefit to the firm of being allowed to increase emissions by that increment.

Relation to Policy Setting

For convenience, we can refer to the function relating emissions and profits, namely Eq. (4), as $\pi(e)$, noting that levels of output and abatement equipment adjust optimally in the background as e changes. Then the MAC (Eq. 5) can be denoted

π_e. If, instead of an emissions constraint of the form in Eq. (3), the firm were charged a tax per unit of emissions, its profit maximising problem would become

$$\max\{w.r.t.\ e\}\ \ \pi(e) - \tau e$$

where τ is the emissions tax rate. The solution is

$$\pi_e = \tau. \tag{6}$$

In other words, the firm chooses an emissions level at which MAC equals the tax rate. In Fig. 1, if the tax rate is m_1, the resulting emissions level would be e_1.

In response to a tradable permits policy, if the competitive price of permits is m_1; the resulting emissions level is e_1, the same as in the tax case. Consequently the MAC corresponds to a firm's demand curve for tradable emission permits. Joskow et al. (1998) used data from the US sulphur dioxide trading market to plot bid–offer curves, which show the number of permits sought at each price level. Since these show the industry's marginal willingness to pay for each additional unit of allowed emissions, they correspond to the MACs of the industry subject to the tradable permits regulation.

The MAC in Eq. (5) is derived by assuming a simple emissions constraint in the form of Eq. (3). If the emissions constraint takes another form, the MAC will also change. In general, the less direct the form of the regulatory constraint, the higher the MAC will be at every emissions level. When the desired policy target is an emissions level, then a single constraint in the form of Eq. (3) yields the minimum MAC. If, for example, the constraint is instead expressed in the form of emissions intensity, or emissions per unit of output:

$$\frac{e(y,a)}{y} \leq z_1 \tag{7}$$

then for the same emissions levels, the MAC must be higher compared to the one associated with Eq. (5) (McKitrick 2010, Chap. 6; Helfand 1991).

The Equimarginal Criterion

For the industry as a whole, operating on the lowest MAC requires that the distribution of emission reduction requirements yields equal marginal abatement cost levels for all individual polluters. Suppose each firm $i = 1, \ldots, n$ faces a constraint in the form of Eq. (3). Then the total profits associated with emissions across all firms is given by the summation $\sum_{i=1}^{n} \pi^i(e_i)$. Minimising the cost of emission reductions is equivalent to maximising this term, subject to a cap on total emissions $E = \sum_{i=1}^{n} e_i$. To solve this we form the Lagrangian function

$$L = \sum_{i=1}^{n} \pi^i(e_i) - \lambda\left(\sum_{i=1}^{n} e_i - E\right) \tag{8}$$

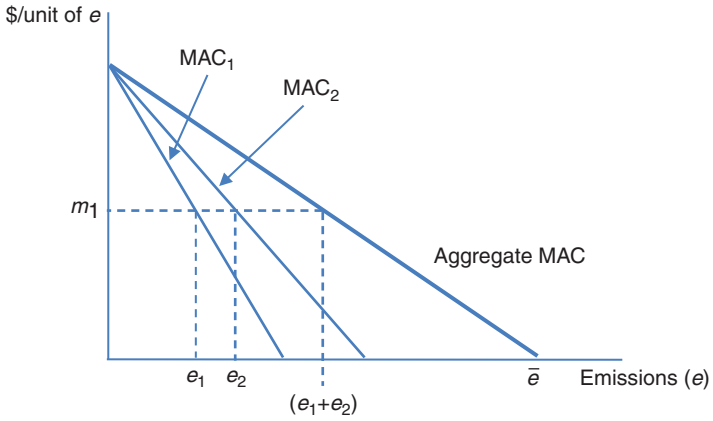
where λ is the Lagrange multiplier. The first-order conditions imply

$$\pi_e^i = \pi_e^j\ \forall\ i \neq j. \tag{9}$$

This is called the equimarginal criterion and states that marginal abatement cost levels must be equal for any arbitrary pair of polluters, and thus across all polluters, in order to yield a cost-minimising distribution of pollution abatement requirements. Since it is typically not the case that regulators have sufficient information to impose firm-specific standards in such a way that Eq. (9) will hold, an alternative is to use a pricing instrument such as an emissions tax τ. Equation (6) applies to all firms, so if all firms face the same tax rate then Eq. (9) must hold.

Aggregation

If the equimarginal criterion holds, then at every price level m_1 the aggregate MAC shows the corresponding total emissions. Figure 1 can be augmented to the two-firm case to illustrate this. As shown in Fig. 3, the aggregate MAC is the horizontal sum of MAC_1 and MAC_2, analogous to the horizontal summation of individual demand curves into a market demand curve.



Marginal Abatement Costs, Fig. 3 Aggregate marginal abatement cost curve

If the equimarginal criterion does not hold, the aggregate MAC is not uniquely defined, since the marginal cost of another unit of emission reductions will depend on which firm or firms abate, and in which order.

Implied Price and Scarcity Rents

Even when quantity is the regulator's target variable, the existence of the MAC implies that there is always a price, or willingness to pay, associated with each level of emissions. Economic instruments such as emission taxes and tradable permits reveal this price, which provides information for a regulator that can be used to refine the implementation towards the optimal level. The price is analogous to (and in case of emission charges, takes the form of) a tax, and like all taxes it generates a marginal excess burden and also exacerbates excess burdens in connected markets. This is called the tax interaction effect, and can be offset in part or in whole by capturing the rents created by the emissions control policy and using them to fund tax reductions elsewhere (Parry 1995).

The interaction effect is still present even when an emissions control policy constrains the quantity rather than imposing a price. Quantity restrictions create scarcity rents that are not captured by the regulator, and therefore are unavailable to fund offsetting reductions in other taxes. Therefore the macroeconomic cost of non revenue-raising policies is higher than revenue-neutral emission pricing policies. This leads to a discontinuity where the MAC meets the emissions axis such that the first unit of emissions reduction is no longer infinitesimally small (Parry et al. 1999).

It is important to note that when emission control policies take the form of constraints on the quantity of emissions, it is incorrect to say that, since there is no emissions tax, economic agents are not 'paying the cost' of their emissions. While they are not remitting an emissions tax to the government, they are paying the cost in the form of higher goods prices and, in some cases, rents that accrue to industries that benefit from quantity restrictions created by the emissions control policy.

Empirical Estimation

A difficulty in estimating the MAC in the form of Eq. (5) is that neither units of abating inputs nor their costs are typically observed, making it infeasible to include them in a cost function framework. Where such data are available, the estimation can be undertaken upon assumption of a specific functional form: for an example, see Yiridoe and Weersink (1998). An alternative approach has been to use output distance functions, which do not require specific data on abatement equipment, instead interpreting input-normalised variations in output as an indicator of overall efficiency, which can indicate costs of

abatement activity. For an example see Kwon and Yun (1999).

Finally, the case of carbon dioxide (CO_2) has received considerable attention, in part because of the connection to the topic of global warming, but also because in almost all cases there are no effective scrubbers or other types of abatement equipment options for reducing CO_2 emissions, which simplifies the empirical problem considerably. Estimation can proceed through direct econometric methods or through the use of computable general equilibrium or other macroeconomic models. For examples see Morris et al. (2012) and the survey in Kuik et al. (2009).

See Also

- ▶ Environmental Economics
- ▶ Lagrange Multipliers
- ▶ Pigouvian Taxes
- ▶ Pollution Permits

Bibliography

Helfand, G.E. 1991. Standards versus standards: The effects of different pollution restrictions. *American Economic Review* 81(3): 622–634.

Joskow, P.L., R. Schmalensee, and E.M. Bailey. 1998. The market for sulfur dioxide emissions. *American Economic Review* 88(4): 669–685.

Kuik, O., L. Brander, and R.S.J. Tol. 2009. Marginal abatement costs of greenhouse gas emissions: A meta-analysis. *Energy Policy* 37: 1395–1403.

Kwon, O.S., and W.-C. Yun. 1999. Estimation of the marginal abatement costs of airborne pollutants in Korea's power generation sector. *Energy Economics* 21: 547–560.

McKitrick, R.R. 1999. A derivation of the marginal abatement cost function. *Journal of Environmental Economics and Management* May: 306–314.

McKitrick, R.R. 2010. *Economic analysis of environmental policy.* Toronto: University of Toronto Press.

Morris, J., S. Paltsev, and J. Reilly. 2012. Marginal abatement costs and marginal welfare costs for greenhouse gas reductions: Results from the EPPA model. *Environmental Modeling and Assessment* 17: 325–336.

Parry, I.W.H. 1995. Pollution taxes and revenue recycling. *Journal of Environmental Economics and Management* 29: S64–S77.

Parry, I., R.C. Williams III, and L.H. Goulder. 1999. When can carbon abatement policies increase welfare? The fundamental role of distorted factor markets. *Journal of Environmental Economics and Management* 37: 52–84.

Yiridoe, E.K., and A. Weersink. 1998. Marginal abatement costs of reducing groundwater-N pollution with intensive and extensive farm management choices. *Agriculture and Resource Economics Review* 27(02): 169–185.

Marginal and Average Cost Pricing

William Vickrey

Abstract

Under perfect competition, marginal cost and average cost of a product are equal to each other and to its price, an arrangement that is Pareto-optimal in the absence of neighbourhood effects. Technical progress is making it possible to vary the prices of some products (such as telephony and electricity) from moment to moment in accordance with marginal cost. Such responsive pricing would help guarantee essential services and reduce the cost of providing reserve capacity. Where there are economies of scale, prices set at marginal cost will fail to cover total costs, thus requiring a subsidy.

Keywords

Airline industry; Average cost pricing; Congestion charges; Depreciation; Economies of scale; Electricity markets; Escrow funds; Excise taxes; Inverse elasticity rule; Land rents; Land tax; Leakage ratio; Long run and short run; Marginal and average cost pricing; Non-price competition; 'One-horse-shay' asset; Price discrimination; Price stability; Quantity–volume interrelationships; Ramsey, F.; Reservation price; Responsive pricing; Second-best pricing; Spatial economics; Sticky prices; Subsidy; Vickrey, W.S.

JEL Classifications

D2

In a pure and simple static world of perfect competition, where production units purchase or rent all their inputs in competitive markets and each sells a single homogeneous product competitively, production takes place at a point of constant returns to scale where the marginal cost and average cost of the product are equal to each other and to its price. If in addition there are no neighbourhood effects or externalities operating outside the market, the result will be Pareto efficient, meaning that there is no feasible alternative arrangement that would be better for someone and no worse for anyone.

Difficulties with the Concept of Average Cost

As soon as production takes place with durable capital facilities that must be adapted to the needs of an individual firm there may no longer be an effective market for these facilities and a cost of their use during any particular period must be determined by other means. In the rather extreme case of the 'one-horse-shay' asset that in a static environment yields a stream of identical services over a known lifetime, a constant periodic rental cost can be derived by the use of a 'sinking fund' method of depreciation in which the rent is the sum of an increasing depreciation charge and a decreasing interest charge on the net value. But where the value of the service varies over time, whether because of physical deterioration, an increasing cost of maintenance needed to keep the item in 'as new' condition, or shifts in demand, this would in principle cause depreciation charges to vary; in practice this is done in one of a number of arbitrary ways by using 'straight-line' or various forms of 'accelerated' depreciation. If these charges are used as a basis for pricing, where competition is imperfect enough to give some leeway, the results can be correspondingly arbitrary.

More serious problems arise in the increasingly widespread cases of joint production of several distinguishable products or services. Where competitive markets exist, the market conditions dictate the allocation of joint costs among the various products, as when a meat-packing establishment produces steaks, hides, glue and offals. There is no way in which one can determine a meaningful average cost of hides by considering only the production process. Where the products, though economically widely different, are physically similar, it is tempting to cut the Gordian knot and average over the entire output, often at the cost of serious impairment of economic efficiency. Even when elaborate rationales are concocted by cost accountants, unless demand conditions as well as production conditions are taken into account the results are essentially arbitrary.

One can do a little better with marginal cost, at least if one is seeking a short-run marginal social cost (hereafter SRMSC), which is the concept that would be relevant for efficiency-promoting pricing decisions. Unless a consumer is presented with a price that correctly represents the marginal social cost associated with the various alternatives open to him, he is likely to make inefficient decisions.

The Importance of Emphasizing the Short Run

One often finds in the literature proposals to use a 'long-run marginal cost' as a basis for setting rates. The trouble is that in an operation producing a multitude of products with interrelated costs it is not possible even to define in any precise way what could be meant by a 'long-run marginal cost', any more than one could define a relevant long-run marginal cost for the hides and steak that are derived from the same carcass in the face of fluctuations over time in relative demand.

The attempt to use a long-run concept seems to be motivated in part by the notion that in some sense the long-run concept is more inclusive in that it allows for variation in capital investment and would include a return on such investment, whereas short-run marginal costs would fail to cover the costs of capital investment. In the single-product steady-state case, however, which is the only case for which the long-run marginal cost can be clearly defined, if the investment in plant is at the optimal level, i.e. the level which

will result in the given output being produced at the lowest total cost, short- and long-run average cost curves will be tangent to each other at the given output, and short- and long-term marginal costs will beequal. Short run marginal-cost prices will therefore cover just as much of the total cost as will prices based on 'long-run marginal cost'. If short-run marginal cost is below the long-run marginal cost, this would indicate that the installed plant is larger than optimum, and conversely if plant is below optimum size, short-run marginal cost will be above long-run marginal cost.

Flexible Versus Stable Prices

A long-run approach is sometimes advocated on the ground that it results in more stable prices. Price rigidity, however, exacts a high toll in terms of reduced efficiency. It is sometimes argued that stable prices are required for intelligent planning for installations that commit the investor to the use of a given volume of service. There is nothing in a SRMSC pricing policy, however, that precludes providing the consumer with estimates of the probable course of prices in the longer term, or even entering into long-term contracts to purchase specified quantities of service. If they are not to interfere with efficiency, however, such contracts should allow for the possibility of purchasing additional amounts at the eventual going rates, or of selling back some of the contracted-for output if this should prove profitable for the consumer.

Lack of flexibility in pricing has, indeed, been a major source of inefficiency in the use of utility services, whether arising as a result of the cumbersomeness of the regulatory procedures in privately owned utilities, or of bureaucratic inertia in publicly owned ones. At times it has even appeared that it takes longer to carry out the bureaucratic procedures involved in altering a price than to install additional capacity, whereas in terms of the underlying capabilities prices can and should be altered on shorter notice than the time taken to adjust fixed capital installations.

Optimal Decision-Making Sequence

The efficient pattern of decision-making consists of first establishing a pricing policy to be followed in the future (as distinct from the application of that policy to produce a specific set of prices), then planning adjustments to fixed capital installations according to a cost-benefit analysis based on predicted demand patterns and predicted application of the pricing policy, subject to whatever financial constraints may be applicable, and then eventually determining prices on a day-to-day or month-to-month basis in terms of conditions as they actually develop.

Too often a rigid adherence to inappropriate financial constraints results in a pattern of pricing over time that leads to gross inefficiency in the utilization of facilities that are added in large increments. In the setting of tolls on bridges, for example, a high fixed toll is often imposed from the start in an attempt to minimize early shortfalls of revenues below interest and amortization charges. When the indebtedness incurred to finance the facility is finally paid off, tolls are often eliminated, sometimes just at the time that they should be increased in order to check the growth of traffic and congestion and defer the necessity for the construction of additional facilities.

The Forward-Looking Character of Marginal Cost

Since changes in present usage cannot affect costs incurred or irrevocably committed to in the past, it is only present and future costs that are of concern in the determination of marginal cost. Past recorded costs are relevant only as predictors of what current and future costs will turn out to be. The marginal cost of ten gallons of gasoline pumped into a car is not determined by what the service station paid for that gasoline, but by the cost expected to be incurred to replace that gasoline at the next delivery. The substantial time-lag that often exists between a change in price at the raw material level and its reflection at the retail level is one of the pervasive failings that

contributes to the inefficiency of the economic system.

Another more important case in which future impacts are of vital importance in the calculation of marginal cost is where congestion accumulates a backlog of demand that has to be worked off over a period of time. A particularly striking case of this occurs when traffic regularly accumulates in a queue during rush hours at a bottleneck such as a toll bridge. The consequence of adding a car to the traffic stream is that there will be one more car waiting in the queue from the time the car joins the queue until the queue is eventually worked off, assuming that the flow through the bottleneck will be unaffected by the lengthening of the queue.

The marginal cost of a vehicle trip will be measured in terms of a number of vehicle hours of delay equal to the interval from the time the car would have arrived at the choke point if there had been no delay, to the time the queue is finally worked off. This is not measured by the length of the queue at the moment, but will be determined by the subsequent arrival of traffic over an extended period. A car arriving at the queue after it began to accumulate at 7:30 may get through the bottleneck at 8:00, after being delayed by only 15 minutes, but if the bottleneck will not be worked off until 10:00 the marginal cost will be 21/4 vehicle hours of which only 1/4 hour is borne by the added caritself. The remaining two hours, if evaluated at $5 per vehicle hour, would indicate that under these conditions the toll that would represent this externality would be $10. Marginal cost cannot be determined exclusively from conditions at the moment, but may well depend, often to an important extent, on predictions as to what the impact of current consumption will be on conditions some distance into the future.

Marginal Cost of Heterogeneous Sets of Uses

It will often happen, for various reasons, that the same price will have to be applied to a non-homogeneous set of uses. To set such a price properly, the marginal costs of the various uses within the set covered must be combined in some way to get a marginal cost relevant to this decision. It would be wrong, however, merely to average the marginal costs of all the uses for which this price is to be charged.

Rather, the decision as to whether a decrease in a given price is desirable must consider the cost of the increments or decrements in the various outputs that will be bought as a result of the price change. In averaging the marginal costs of the various usage categories, the weighting will have to be in proportion to the responsiveness of each usage category to the change in price.

For example, if a price is to be set for electricity consumption on summer weekday afternoons, in a system where air-conditioning is an important load, consumption and marginal cost may be higher on hot days than on warm days, but it may be considered too difficult to differentiate in price between the two categories of days. An increase in the price for this entire set of periods may induce some customers to adjust the thermostat setting. But during hot days the equipment may work full tilt without reducing the temperature to the thermostat setting, whereas on warm days there will be a reduction in power consumption. The marginal cost relevant to the setting of the common price would then be determined predominantly by the lower marginal cost of the warm-day consumption, and relatively little, if at all, by the higher marginal cost hot-day consumption.

Anticipatory Marginal Cost

In many cases a customer will make his effective decision to consume an item some time in advance, and it will be the expected price as perceived by him at that time that determines his decision. If, as in services subject to reservation, a firm price must be quoted the time the reservation is made, it is the expected marginal cost as of that moment that should govern the price charged. In the case of a service where the demand is highly variable and to a considerable extent unpredictable, such an expected marginal cost would be an average of marginal costs that might arise under alternative possible

developments, possibly ranging from a very low value, if there turns out to be unused capacity, to the possibly quite high value if another latecomer must be turned away. The respective probabilities of these outcomes, as estimated at a given time, will vary with the proportion of the total supply already sold, the time remaining to the delivery of the service, and the pricing policy to be followed in the interim.

At one extreme, for long-haul airline reservations where the unit of sale is large, one might find it worth while to have a fairly elaborate pricing scheme in which the price quoted would vary according to the proportion of seats on a given flight already sold and the time remaining to departure, in simulation of what an ideal speculators' market might produce, the price at any time being an estimate of the price which, if maintained thereafter, would result in all the remaining seats being just sold out at departure time. This would correspond to marginal cost in that the sale of a seat at any given time would slightly raise the price during the remaining period to decrease demand by one unit, at a price that would be expected to be on the average equal to the price at which the seat was sold, indicating that the price was equal to the value of the seat to the alternative passenger.

Quality-Volume Interrelationships

In principle, in the absence of barriers to entry, competition would induce the supply of just sufficient seats on the various routes to cause revenues produced by such pricing to just cover costs. Even this, however, would be optimal only on those routes where traffic is so heavy that even with planes of a size producing the lowest cost per seat, further increases in service frequency would be of negligible value. On most routes there will remain economies of scale in that either providing more seats at the same frequency of service with larger planes would reduce costs per seat, or providing more seats with the same size of planes would provide an increased frequency of service that would be of value to others than the additional riders. In the latter case the marginal cost of providing for the additional passengers would be calculated by deducting the increase in the value of the service by reason of increased frequency from the cost of providing the added seats.

If it were possible to adjust plane size and frequency in a continuous fashion, then if the situation is optimal the two marginal costs would be equal. In practice both plane size and service frequency can be varied only in discrete jumps, so that this relation would be only approximate. Optimal price would be above a downward marginal cost calculated on the basis of a reduction in service, and below an upward marginal cost calculated on the basis of an increase in service. The decreases and increases might involve a combination of frequency and plane size changes. To preserve the formulation that price should equal marginal cost it may be useful to define marginal cost in such cases as consisting of the range between these upward and downward values rather than as a single point.

In practice, between the existence of economies of scale and the imperfect cross-elasticity of demand between flights at various times and with different amenities, removal of regulation tends to result in an emphasis on non-price competition, attempts to subdivide the market by various devices and restrictions to permit discriminatory pricing, and a bunching of service schedules at salient times and places that provides a lower overall level of convenience than would be possible were the given number of seat miles distributed more efficiently.

Where the unit of sale is small it may not be worth while to incur the transaction costs of varying price in strict conformity with SRMSC. One could, in theory, apply the same principle to the sale of newspapers at a given outlet. The price of a newspaper would vary according to the number of unsold papers remaining and the time of day. This would result in less disappointment of customers having an urgent desire for a paper late in the day and encountering a sold-out condition, and fewer unsold papers returned. But unless some ingenious device can be found for executing such a programme at low transaction costs, it probably would not be considered worth while, even by the most sanguine advocate of marginal cost pricing.

Wear and Tear, Depreciation and Marginal Cost

Even in the absence of lumpiness or technological change, existing methods of charging for capital use often fail to give a proper evaluation of marginal cost. This is especially true where the useful life of a unit of equipment is determined more by amount of use than by lapse of time. In the extreme case of equipment that must be retired at the end of a given number of miles or hours of active service, or after the production of so many kwh of energy, and which, in one-horse-shay fashion, gives a uniform quality of service over its lifetime without requiring increasing levels of maintenance, the marginal cost of use at a given time will be the consequent advancing of the time of retirement of the equipment. The marginal cost of using the newest units will be the lowest, and will advance over time at a rate equal to the rate of interest as the equipment ages and the advancement of replacement consequent upon use becomes less and less remote.

In a service subject to daily and weekly peaks, the newest equipment will be allocated to the heaviest service, operating during both peak and off-peak hours. Equipment will be relegated to less and less intense service as it ages. The marginal cost of service at a particular moment will be that for the oldest unit that has to be pressed into service at that instant. The rental charge for the use of the unit will vary gradually over the entire range of demands, rather than dropping off to zero whenever the full complement of equipment is not required. At the other end, in this extreme case, the service provided would not necessarily be held constant by price variation over an extended peak period: under the conditions postulated it would be possible to provide for needle peaks by planning for the stretching out over time of the final service units of the oldest equipment. In this way the required peak capacity can be provided at a cost much lower than that which would be calculated by loading all the capital charges for the added equipment on this brief period of use.

Another way of looking at the matter is to appeal to the proposition that perfect competition under conditions of perfect foresight will produce optimal results. To this end one can suppose a situation in which vehicles are rented by the hour from a large number of lessors operating in a competitive market. For simplicity, initially, one can assume all vehicles to be of the one-horse-shay variety, being equivalent to bundles of hours of active service, with the quality of service being independent of age up to a final 'bubble-burst' collapse. Also, for simplicity, assume a steady state in which vehicles are scrapped and replaced at a constant rate over time, so that at any given moment vehicles are evenly distributed by age.

A common market rental price for all vehicles at any given time of the week will emerge, being higher as the number of vehicles in service at the time is greater. During any given week, each renter will have a reservation price for his vehicle, such that he will rent his vehicle during those hours for which the market rental is above this reservation price and never when the market rate is lower. This reservation price will increase over time for any given vehicle at the market rate of interest, since a renter will rent his vehicle if and only if the net present value of the rental discounted back to the time of purchase exceeds some fixed amount. The owner would not want to rent his vehicle for a net present value less than he could have got by selling one of this stock of service units at some other time at or just below his reservation price. New buses will have the lowest reservation price and will be assigned to the schedules calling for the most hours of service per week, while old buses will be held idle during slack hours and used only for peak service. As each bus ages it will be assigned to less and less heavy service along the load-duration curve.

This pattern of usage can be regarded as resulting from a desire to recover the capital tied up in the usage units of each bus as rapidly as possible. It is related to the practice in electric utilities of using the newest units for peak service, in that case motivated in part by the tendency for the newer units to be more efficient in thermal terms. To be sure, occasionally new units are designed specifically for peaking service, with a correspondingly low capital cost, though this is a relatively recent phenomenon related to a

slowing-down of secular increases in potential thermal efficiency.

In any case, where wear-and-tear is a factor, one cannot properly allocate depreciation charges primarily to peak service, however defined, nor should they be spread evenly over all service, much less spread evenly over hours of the week so that vehicle hours in off-peak periods would get higher charges than during the peak. Rather the depreciation charge per vehicle hour will vary gradually and in a positive direction with the intensity of use of the equipment at any given hour.

The analysis becomes a little complicated when equipment life is dependent on mileage or loading or intensity of use as well as hours of active service, so that different rentals would properly be chargeable according to the nature of the service for which the unit is being rented. Also further analysis is required if equipment is laid up between runs at isolated terminals rather than at a central depot where a market could be postulated, or if the fleet contains vehicles varying in size or other characteristics. It would even be theoretically appropriate to charge different fares for the same trip at the same time if made on vehicles with different origins or destinations. (In Hong Kong, indeed, the practice is to charge a flat fare on each route, but to differentiate the fare fairly elaborately as among routes. On segments where routes converge, this has the unfortunate result of unduly concentrating riding on buses with the lower fares, even where the higher fare buses have empty seats and are making stops in any case for other passengers.)

Costs of major overhauls that are performed at relatively long intervals would also complicate the picture. There are also problems associated with gradual or sudden changes in overall demand levels, or special events that can be anticipated sufficiently to present an opportunity for reacting in terms of a change in price. The picture can be further complicated if, as was discussed above, there are changes in available technologies or other changes in quality or cost. But the same method of analysis in terms of a hypothetical competitive market can be used to obtain appropriate results.

For the sake of simplicity the above analysis has been couched mainly in terms of a bus service, but the analysis is applicable wherever the useful life of equipment is in part a function of the intensity with which it is used.

Responsive Pricing

In some cases, notably in telephone and electric power services, the technical possibility exists for conveying information as to the current price to customers at the instant of consumption, and for customers to respond to such information in a worthwhile manner at modest cost. In the case of telephone service the information as to the level of charges for local calls can be substituted for the dial tone, with information on rates for long distance calls provided to users who wait for it before dialing the final digits. If the charge exceeds what the customer is willing to pay the call can be aborted with little occupancy of equipment or inconvenience to the user. Prices can be varied from moment to moment in accordance with marginal cost, as estimated from the degree of busyness of the relevant sets of equipment.

In the case of electric power, the costs of providing for a variation of the price according to the conditions of the moment would be somewhat greater. But if the facilities take the form of remote meter reading, either by carrier current over the power lines or by a separate communications channel, much of the cost would be covered by the avoidance of costs involved in manual meter reading. A signal of rate changes can then be provided to the customer as a by-product of the signal required to initiate a new rate period. The customer can then respond either manually or by installing automatic equipment which will adjust the operation of such items as air-conditioning and refrigeration compressors, water heaters and the like, according to the level of rates in a manner determined by the customer himself. Retrofitting of existing meters by attaching a pulse-generating device such as a mirror and photo-electric cell to the rotor shaft of the existing meter and feeding the pulses to electronic counters and registers should be possible at relatively low cost.

Such responsive pricing would be especially valuable in dealing with emergencies, providing greater assurance of the maintenance of essential services than is possible with existing techniques, and making it possible to reduce substantially the cost of providing reserve capacity. In the case of floods, conflagrations, breakdowns in transit, or other emergencies that under present conditions tend to result in the overloading of telephone facilities and difficulties in completing calls of a vital nature, rates can be charged that are high enough to inhibit a sufficient number of less important calls so that the ability of the system to handle vital calls promptly is preserved. This is difficult to do with present techniques, for while it is relatively easy to give priority to calls originating at such points as police stations, hospitals, and the like, most emergency calls are calls to rather than from these points and it is much more difficult to distinguish such calls close to the point of origin. And there are always a certain number of vital calls not distinguishable in terms of either origin or destination.

Again, in the case of unscheduled power cuts, it would be possible to cause an almost instantaneous shedding of substantial water-heating and refrigeration loads, followed, in the case of an extended cut, by partial shedding of elevator, transit and batch process loads for which it is more inconvenient to respond quite so promptly, after which a sufficient refrigeration load can be picked up as needed to avoid food spoilage. Many of the serious consequences of major power blackouts could have been avoided had such a system been in place at the time. Reserve capacity might well be cut back to provision for scheduled maintenance, leaving the load-shedding capability of responsive pricing to function as a reserve. In many cases the speed of response possible with responsive pricing would be faster than the reaction time within which reserve capacity can pick up load, leading to better voltage regulation and a higher quality of service to customers remaining on the line. And if, in spite of everything, areas must be cut off completely, responsive pricing would also be of considerable help in facilitating a smooth recovery from an outage: instead of having a whole army of motors trying to start up at once upon the restoration of power, with consequent load surges, voltage fluctuations, and malfunction of equipment, load could be picked up smoothly and gradually as the price is lowered from the inhibiting level.

Preserving Incentives with Escrowfunds

With privately owned utilities the regulatory process is too slow to permit prices established directly by regulation to be constantly adjusted to changing current conditions, unless indeed the regulators were to assume a large part of what are normally the responsibilities of management. The problem thus arises of how to allow the prices to be paid by customers to be varied by the utility management without giving rise to incentives for behaviour contrary to the public interest. Even if a formula could be devised that would require the utility to adjust prices to track short-run marginal cost, if the utility were allowed to keep the revenues thus generated without restriction, this would set up undesirable incentives for the utility to skimp on the provision of capacity in order to drive up the marginal cost, price, revenues, and profits.

A resolution of this dilemma can be achieved by separating the revenue to be retained by the utility from the amounts to be paid by customers. We can have the 'responsive' prices paid by customers vary according to short-run marginal social cost, while the revenues to be retained by the utility are determined by a 'standard' price schedule fixed by regulation in the normal manner, the difference being paid into or out of an escrow fund. Failure of the utility to expand capacity adequately would drive marginal cost up, and with it the responsive price, causing revenues to flow into the escrow fund, but the only way the utility could draw on these funds would be to expand capacity sufficiently to drive marginal cost down, causing the responsive rate to fall below the standard rate on the average, entitling the utility to make up the difference from the escrow fund as long as it lasts. Excessive expansion would result in the escrow fund being exhausted, with a corresponding constraint on

the revenues obtainable by the utility from the unaugmented low responsive rates.

The setting of the responsive rates would have to be to a large extent at the discretion of the operating utility, though the regulatory commission could monitor the process and even attempt to establish guidelines according to which the responsive price should be set. The utility would normally have no incentive to set the responsive rate below marginal cost, since this would merely increase sales and hence costs by more than any possible long-run increase in revenues to the utility. To be sure, in the short run it might be able to draw on the escrow fund to the extent of the excess, if any, of the standard rate over the responsive rate, but since from a long-run perspective there will normally be other more advantageous ways of drawing on this fund this will not be attractive.

When marginal cost is below the standard price, which would tend to be the usual situation, the utility would in general have an incentive to set the price between the marginal cost and the standard price, since each additional sale produced by the lower price will yield an immediate net revenue equal to the difference between marginal cost and the standard price, offset only by the drawing down of the escrow fund by the difference between the responsive and the standard price. When marginal cost is above the standard price, which with a properly designed standard rate schedule with time-of-day variation should happen relatively rarely, the utility would have an incentive to set the price at least at the marginal cost level, since to set it lower would tend to increase output at a cost in excess of anything the utility could ever recover. How much higher than marginal cost the price might be set would in theory be limited by the condition that the price could not be high enough to curtail demand sufficiently to drive marginal cost below the standard price. If the standard price has an adequate time-of-day variation, this constraint, loose as it may seem, may be sufficient. Additional guidelines could of course be imposed by the regulatory commission for those rare occasions where this constraint might seem insufficient to keep prices within bounds.

Actual Steps Towards Responsive Pricing

Some actual practices of utility companies are steps in the direction of responsive pricing. Contracts for ‘interruptible’ power provide for load shedding at the discretion of the utility subject to some overall limits. As these are fairly long-term contracts that usually require ad hoc communication between the utility and the customer, their applicability is limited and there is no assurance that the necessary shedding will be done in the most economical manner. Many customers are reluctant to submit to load shedding that is not under their control at least to some extent, and that might be imposed under awkward circumstances. Where reserves are ample and interruption is highly unlikely, such contracts have been challenged as being a form of concealed discriminatory concession. On the other hand customers entering into such contracts in the expectation of not being interrupted may feel aggrieved if interruption actually takes place.

Another experimental provision applied by a company with a heavy summer air-conditioning load is for a special surcharge to be applied to the usage of larger customers on days when the temperature at some standard location exceeds a critical level. And another company bases its demand charge on the individual customer’s demand recorded at the time that turns out to have been the monthly system peak load, supplying the customers with information as to moment-to-moment variations in the system load. This leads to interesting game-playing on the part of customers as they attempt to keep their own consumption down at times that look as though they might become the monthly peak, with the result that this action may itself shift the peak to another time.

Economies of Scale, Subsidy and Second Best Pricing

Where there are economies of scale, prices set at marginal cost will fail to cover total costs, thus requiring a subsidy. One reason for wanting to avoid such a subsidy is that if an agency is

considered eligible for a subsidy much of the pressure on management to operate efficiently will be lost and management effort will be diverted from controlling costs to pleading for an enhancement of the subsidy. This effect can be minimized by establishing the base for the subsidy in a manner as little susceptible as possible to untoward pressure from management. But it is unlikely that this can be as effective in preserving incentives for cost containment as a requirement that the operation be financially self-sustaining. To achieve this, prices must be raised above marginal cost, and in a multi-product operation the question arises as to how these margins should vary from one price to another within the agency.

Another objection to subsidy is that it raises hard questions of who should bear the burden of the subsidy. More fundamentally the taxes imposed to provide the subsidy will often have distorting effects of their own, and minimizing the overall distortion would again require prices to be raised above marginal cost. One can, indeed, regard these excesses of price over marginal cost as excise taxes comparable to other excise taxes that might be levied to raise a specified amount of revenue.

The answer given to the problem of how to allocate excise taxes and other margins of price above marginal cost so as to minimize the overall loss of economic efficiency given by Frank Ramsey in 1927 can be expressed for the case of independent demands as the inverse elasticity rule, which says that the margin of price over marginal cost as a percentage of the price shall be inversely proportional to the elasticity of demand. A more general formulation is one that states that prices shall be such that consumption of the various services would be decreased by a uniform percentage from that which would have been consumed if price had been set at marginal cost and demand had been a linear extrapolation from the neighbourhood of the 'second-best' point.

A more transparent formulation, devised by Bernard Sobin in work for the US Postal Service, is the requirement of a uniform 'leakage ratio', leakage being the difference between the net revenue actually derivable from a small increment in a particular price and the hypothetical revenue that would have been obtained had there been no change in consumption as a result of this increment. Leakage is the algebraic sum of the products of the changes in consumption of the various related products induced by the small change in a given price, and the respective margins between their prices and marginal costs. Leakage is a measure of the loss of efficiency resulting from the change in the particular price, and the leakage ratio is the ratio of this loss of efficiency to the hypothetical gain in gross revenue if there had been no change in consumption. If one leakage ratio should be greater than another, the same net revenue could be obtained at greater economic efficiency by getting more revenue from the price with the smaller leakage ratio and less from the other. The second-best solution accordingly requires that all leakage ratios be equal.

This analysis can be extended to the case where the agency is being subsidized by taxes which involve an adverse impact on the economy, in terms of marginal distorting effects, compliance costs, and collection costs, which can be expressed as the 'marginal cost of public funds' (MCPF). For a net decrease in the subsidy derived by increasing a price, which can be considered to be equivalent to imposing a tax equal to the difference between the marginal cost and the price, $\mathrm{MCPF} = \mathrm{LR}/(1 - \mathrm{LR})$, where LR is the leakage ratio. A second-best optimum is then one where the MCPF's are equalized over both external and internal taxes.

Special Sources of Subsidy: Land Rents and Congestion Charges

In the case of goods and services with economies of scale that are provided primarily to consumers within a particular urbanized area, methods of financing may be available that involve no marginal cost of public funds or even result in an enhancement of efficiency. The existence of large cities, indeed, is to a predominant extent due to the availability in the city of goods and services produced under conditions of economies of scale: if there were no economies of scale,

activity could be scattered about the landscape in hamlets, with great reduction in the high transportation costs involved in movement about a large city. If prices of these services are reduced to marginal cost, the increased attractiveness of the city as a consequence would tend to drive up land rents within the city, and it appears quite appropriate that a levy on such rents should be used to finance the required subsidies. And while there are practical and conceptual difficulties in defining exactly how land rents or land values should be specified for purposes of levying a tax, it is generally considered that a tax on land values, properly defined, has negligible adverse impacts on the efficient allocation of resources.

Indeed, there is a theorem of spatial economics which states that in a system of perfect competition among cities, the availability in the city of services and products subject to economies of scale, priced at their respective marginal social costs, will generate land rents just sufficient to supply the subsidies required to permit prices to be lowered to marginal cost. Among the more important of these services are utility services such as electric power, telephone, cable communications, water supply, mail collection and delivery, sewers and waste disposal, and local transit. It is not clear just how broad the conditions are under which this theorem would hold, and there are difficulties in capturing all land rents for subsidy purposes, but steps in this direction are clearly desirable.

On a more intuitive level, one can note that a person who occupies or uses land that is provided with services such as the availability of transit, electricity, telephone, mail delivery and the like will be requiring that these services be carried past his property to serve others whether or not he himself uses them. The user of tennis courts located conveniently in a built-up area should no more be excused from contributing to the costs of carrying these services past the courts, even though no direct use is made of electric power, telephone, mail, or other services, than he should expect his auto dealer to cut the price of an automobile by the cost of the headlights and windshield wipers merely because he asserts that he will never drive at night or in bad weather. Tennis players will indeed pay a rent enhanced by the presence of these services and the consequent greater demand for the land for other purposes, but the rent will go to the landlord, not to the purveyor of the services, and the price of the services to those who do use them will be too high for efficiency, unless indeed they are subsidized by other taxes that have their own distorting effects.

It is a corollary of this theorem that it would be to the advantage of the landlords in the area, *faute de mieux*, to agree collectively to pay a tax based on their land values, in order to subsidize the various utility services to enable the prices to be set closer to marginal social cost. They could expect in the long run that this action would increase their rents by as much or more than the taxes. To be sure, they might do better by getting someone else to pull their chestnuts out of the fire, but they can do this only at considerable damage to the overall efficiency of the economy of the city, to say nothing of the inequity of such a parasitic relationship.

In addition to land rents in the conventional sense, there is the land used for city streets for the use of which no adequate rental is generally charged. Charging on the basis of SRMSC for the use of congested city streets would in most cases yield a revenue far in excess of the cost of maintaining such facilities, which could appropriately be used for the subsidy of other urban facilities. Properly adjusted, such charges would increase efficiency by bringing home to the users the costs that their use directly imposes on others.

Formerly it would have been considered impractical to attempt to charge for the use of city streets according to the amount of congestion caused: the collection of tolls by manual methods at a multitude of points within the city might well create more congestion than it averted. Advances in technology have, however, made it possible to do this at minimal interference with traffic flow and at modest cost. One method, proposed as long ago as 1959 and recently carried to the point, is to require all vehicles using the congested facilities to be equipped with electronic response units which will permit individual vehicles to be identified as they pass scanning stations suitably

distributed within and around the congested area so that the records thus generated can be processed by computer and appropriate bills sent to the registered owners at convenient intervals. If properly done, this would greatly improve traffic conditions so that the net cost of the revenue to the road users would be far less than the amount collected as revenue. A pilot installation has recently been tested in Hong Kong with satisfactory results, but full implementation appears to have been deferred, because of the political situation associated with the impending transfer of sovereignty.

Indeed, one can define 'hypercongestion' as a condition where so many cars are attempting to move in a given area that fewer vehicle miles of travel are being accomplished than could be if fewer vehicles were in the area but could move more rapidly; for example if 1000 vehicles in an area move at 8 mph and produce 8000 vehicle miles of travel per hour, reducing the number of cars in the area at a given time to 800 might raise speed to 11 mph producing 8800 vehicle miles of travel per hour. By restricting the flow of traffic in the period leading up to the hypercongestion period, road pricing could prevent hypercongestion from occurring, except possibly sporadically, and in any case so improve conditions that more movement would be accomplished during the peak period at faster speeds. The improvement during peak periods might even be such that total movement throughout the day would be increased, and where conditions are now severe users could find that they are better off than before, even inclusive of the payment of the congestion charge.

If there are bridges, tunnels, or other special facilities for which a toll is already being charged, and which regularly back up a queue during the morning rush-hour, substantial revenues can be obtained at no overall net cost to the users by adding a surcharge to the toll during the period where queueing regularly threatens, rising gradually from zero to a maximum and down again in such a way that by gradual adjustment regular queueing is substantially eliminated. The toll surcharge will then be taking the place of the queue in influencing decisions as to when to travel, and in general those who plan their trips in terms of time of arrival at their destination will be able to leave as many minutes later as they formerly wasted in the queue, pass the bottleneck at the same time as before, and arrive at their destinations at the same time as before. The extra toll will be roughly the equivalent of the value of the extra time enjoyed at the origin point, and the revenue will in effect be obtained at no net burden on the users. In practice the results may be even better than this as a result of the added encouragement to car-pooling, the reduction of obstruction to cross-traffic, and the expediting of emergency or other trips where the delay had been a particularly serious matter.

Gains in the evening may be not quite so dramatic. The situation is not symmetrical, as typically the timing of the trip will be determined in terms of time of departure, which is separated by the queue from the time at the bottleneck. On the other hand the risk of conditions approaching gridlock is greater, since the accumulation of queues inside circumferential bottlenecks is more likely to create congestion, and there is less of a physical barrier to the simultaneous emergence of large quantities of traffic from parking lots into the downtown streets than there is in the morning to the convergence on the congested area of traffic arriving from the outside.

Congestion charges should be imposed, at least notionally, without exception on all forms of traffic. Such charges would be a necessary element in the cost-benefit analysis by which decisions are made as to the level and pattern of bus service to be provided, even though they would not be directly relevant to the determination of the price structure to be applied to that service.

Paradoxes in the Behaviour of Marginal Social Cost

A strict calculation of marginal social cost in particular circumstances may produce what may appear to be quite paradoxical results. For example, in many circumstances it will be optimal, and even essential, to maintain at least a minimum frequency of service in off-peak hours with buses of a standard size, resulting in there being practically always a large number of empty seats

in each bus. Under these circumstances the cost of carrying additional passengers is predominantly the cost of boarding and alighting, including the time of the driver and the other passengers on the bus who are delayed in the process. This cost will be relatively higher if the bus is half full than if it is nearly empty. The result is likely to be that the cost of a trip from a point near one end of the run to a point near the other end, at both of which points the bus is likely to be lightly loaded, may be smaller than for a shorter trip between points near the middle of the run where the bus is likely to be more heavily loaded. This is not a trivial matter: if it were there would be no sense to the refusal of express buses with empty seats to pick up local passengers. It is highly unlikely, however, that fares based on such a seemingly perverse behaviour of cost would meet with popular approval. Indeed, the original US interstate commerce legislation contained prohibitions against higher rates being charged shorter hauls than for any longer hauls within which they might be included.

Another paradoxical example can occur in mixed hydrothermal electric power systems: an increase in fuel prices could result in the marginal cost of power at particular times being reduced rather than increased. If hydro dams are spilling water at certain seasons of the year, increased fuel costs may make it economical to increase the installed generating capacity to make use of the spilling water, even for a briefer period of time over the year than was previously worth while. If during the wet season installed hydro generating capacity is more than sufficient to meet trough demand, marginal cost during such periods will be substantially zero, or at most limited to a small element of wear and tear on equipment pressed into service.

Installing more turbo-generators would expand the period during which this low marginal cost is effective, so that while increased fuel costs cause marginal cost to rise during the peak, the result could also be to lower marginal cost in these intervals into which the period of exclusive hydro supply expands.

In the case of long distance telephone service, the drastic reductions in the cost of bulk line-haul transmission have created a situation where distance, especially beyond the range where separate wire transmission is economical, is relatively unimportant as a cost factor, and where satellite transmission is involved, ground distance is indeed irrelevant. What remains important is the number of successive circuits, with their associated termination and switching equipment, involved in the making of a call. Thus a call between two small communities over a moderate distance, for which the volume of calling is insufficient to warrant the provision of a separate circuit, will generally cost substantially more than a call between important centres over a much longer distance, since the latter will involve only a single long-haul circuit, while the former will require patching through two or more long-haul circuits.

Another anomaly occurs when an innovation promising substantial reductions in costs appears on the horizon, such as has happened repeatedly in telecommunications. Any further installation of the old technology in the interim before the new technology is actually available will involve an investment which will have its capital value diminished over a brief period to that determined by its competition with the new technology. High depreciation or obsolescence charges are in order, and the prospect of the new lower costs results in higher current prices which would serve to hold back current demand and lessen the amount of old technology required to be installed.

Marginal cost pricing is thus not a matter of merely lowering the general level of prices with the aid of a subsidy; with or without subsidy it calls for drastic restructuring of pricing practices, with opportunities for very substantial improvements in efficiency at critical points.

See Also

- ▶ Congestion
- ▶ Ideal Output

Bibliography

Beckwith, B.P. 1955. *Marginal Cost Price-Output Control*. New York: Columbia University Press.

Mitchell, M., G. Manning, and J.P. Acton. 1978. *Peak-load pricing*. Cambridge, MA: Ballinger.
Nelson, J.R., ed. 1964. *Marginal Cost Pricing in Practice*. Englewood Cliffs: Prentice-Hall.
Ramsey, F. 1927. A contribution to the theory of taxation. *Economic Journal* 37: 47–61.
Vickrey, W. 1967. Optimization of traffic and facilities. *Journal of Transport Economics and Policy* 1 (2): 1–14.
Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review* 59: 251–260.
Vickrey, W. 1970. The city as a firm. *The economics of public services*, Proceedings of a conference held by the International Economic Association, Turin, Italy, ed. M.S. Feldstein and R.F. Inman, 334–343. London: Macmillan; New York: Wiley.
Vickrey, W. 1971. Responsive pricing of public utility services. *Bell Journal of Economics and Management Science* 2 (1): 337–346.

Marginal Efficiency of Capital

John Eatwell

The variety of attempts to generate neoclassical results in a 'Keynesian' framework, and 'Keynesian' results in a neoclassical framework, together point to important failings in the *General Theory*. I will argue that the key failures are the inadequacy of Keynes's critique of the neoclassical theory of output and the important ambiguities introduced into the analysis by his marginalist treatment of the labour market and by his portrayal of the marginal efficiency of capital as an elastic demand schedule for investment. Garegnani (1978, 1979) has argued that these failings may be remedied by application of the results of the debate on the neoclassical theory of capital derived from Sraffa's *Production of Commodities*. I will illustrate this point by reference to the implications of the debate for Fisher's analysis of investment and the rate of interest which Keynes identified with his own analysis.

Keynes defined the marginal efficiency of capital as follows:

> If there is an increased investment in any given type of capital during any period of time, the marginal efficiency of that type of capital will diminish as the investment in it is increased, partly because the prospective yield will fall as the supply of that type of capital is increased, and partly because, as a rule, pressure on the facilities for producing that type of capital will cause its supply price to increase; the second of these factors being usually the more important in producing equilibrium in the short run, but the longer the period in view the more does the first factor takes its place. Thus, for each type of capital we can build up a schedule, showing by how much investment in it will have to increase within the period, in order that its marginal efficiency should fall to any given figure. We can then aggregate these schedules for all the types of capital, so as to provide a schedule relating the rate of aggregate investment to the corresponding marginal efficiency of capital in general which that rate of investment will establish. We shall call this the investment demand-schedule; or, alternatively, the schedule of the marginal efficiency of capital (Keynes 1936, p. 136).

Keynes's argument is more complicated than may at first appear, involving as it does assumptions on both the supply and demand conditions for individual capital goods in both short and long run and, finally, at both individual and aggregate levels – the ultimate objective being the derivation of the relationship between the 'rate of aggregate investment' and 'the corresponding marginal efficiency of capital in general', or, to put it another way, the general rate of return.

Taking first the short-period aspect of the argument, Keynes's assumption that increased investment in a given type of capital good will lead to higher cost of production – rising supply price – is quite unfounded. Any short-period situation, and particularly a short period in which capital capacity is widely underutilized, will be characterised by *excess* stocks of materials and machines in some (maybe all) sectors, with (perhaps) shortages in a few sectors too. In such a situation no definite hypothesis may be made as to the likely effect of increased output on cost, though in

conditions of widespread excess capacity it seems reasonable to suppose that costs will tend to *fall* as fixed costs are averaged over higher output. 'Pressure on the facilities' for producing a given capital good will only tend to become significant as full employment is approached, and even then the consequences for the cost of production of an increase in supply of any one capital good cannot be predicted with confidence.

The short-run influence of the demands for capital goods on 'prospective yield' to be derived from further investment are likewise unpredictable, and as to the aggregate effect of all this – nothing can be said at all. Indeed, there is no short-run 'marginal efficiency of capital *in general*' to say anything about! The relationship which Keynes sought must be a long-run relationship, in the sense that it is sufficiently unambiguous and persistent to allow definite conclusions to be drawn concerning the influence of a given volume of investment on the rate of return.

Now, in the longer run Keynes himself suggested that increased output will not result in any increase in cost. Any diminution in return must, therefore, derive from the fall in prospective yield as more capital goods compete to sell their services. What then is the relationship between the volume of investment and the rate of return in the longer run, that is in a situation in which the cost minimising combination of factors is chosen? At the partial level Keynes first considers, the answer seems clear: if all other prices in the economy are taken as given, then *ceteris paribus* it may be argued that there is an inverse relationship between the rate of return and the quantity of capital invested in the production of a given output. But Keynes's argument is on very shaky ground when he attempts to define the relationship for the economy as a whole by simple aggregation of these partial effects, for he can no longer use the *ceteris paribus* condition to keep at bay some fundamental problems.

These fundamental difficulties in Keynes's characterization of the marginal efficiency of capital may be clarified by returning to Fisher's analysis of the incremental rate of return on investment which Keynes tells us is 'identical with my definition' (Keynes 1936, p. 140).

Fisher's analysis is based on the substitution of capital for labour in a full-employment equilibrium, and throughout his discussion of the theory of saving, investment and interest, he imposes a major limitation on his argument – he assumes that all prices, wages and rents are fixed, and do not vary with variations in the rate of interest (Fisher 1930, p. 131n). This 'fixedprice' assumption allows Fisher to express all magnitudes in terms of 'money', and to move between discussion of individual behaviour and that of the economy as a whole without considering the inter-relationship between the rate of interest and prices.

An attempt to generalize Fisher's analysis to a many-commodity model, and hence to relate the determination of prices to the determination of the rate of interest, has been made by Solow (1963, 1967). I have analysed Solow's model and the debate it provoked elsewhere (Eatwell 1976); for our purposes we need only summarize my main conclusions.

It is assumed by Solow that the economy is in a stationary state, producing a consumption good, corn, by means of many reproducible inputs and labour. To enable the definition of limits we may further assume that the technical possibilities of the economy are characterised by a wage-profit frontier which is an envelope to an infinity of wage-profit curves, such that techniques are arrayed continuously along the frontier. Furthermore, consumption and value of capital per head associated with the variation in technique may be described by differentiable functions.

Since the techniques used in the production of corn require inputs of commodities other than corn, the wage-profit line for each technique may assume any negatively sloped curvature. But consumption good output per capita, c, (ie net output per head) and the value of produced inputs per capita, k, are continuous differentiable functions of the rate of interest (rate of profit), r, even though the technique in use varies continuously with r:

$$c = z(r) \quad k = \frac{\text{net output} - \text{wages}}{\text{rate of profit}} = [z(r) - g(r)]/r \tag{1}$$

where $g(r) = w$ is the equation of the wage-profit frontier.

The rate of return over cost of a transition between the technique in use at r and the 'adjacent' technique at $r + h$ is the ratio of the value of the difference in the perpetual consumption streams to the value of the difference in the capital stocks (i.e. the sacrifice required to effect the transition):

$$[z(r+h) - z(r)] / \left[\frac{z(r+h) - g(r+h)}{r+h} - \frac{z(r) - g(r)}{r}\right] \quad (2)$$

'In the limit, as the number of techniques grows denser', $h \to 0$ and expression (2) becomes:

$$z'(r)\frac{r^2}{r[z'(r) - g'(r)] - z(r) + g(r)} \neq r; \quad (3)$$

the marginal rate of return over cost is not equal to the rate of profit. The equality would hold iff:

$$z(r) = g(r) - rg'(r) \quad (4)$$

This would be the case of an economy having the properties of Samuelson's (1962) surrogate production function, and would indicate that, to all intents and purposes, the economy under consideration was set in a one-commodity world. The inequality does not depend on the presence of reswitching or even perversity. So long as the economy contains more than one produced input the rate of profit is not equal to the rate of return over cost. Or, more generally, no demand schedule for investment as a function of the rate of interest may be constructed.

The lack of any logical foundation for the construction of an elastic demand schedule for investment as a function of the rate of interest is simultaneously a critique of the neoclassical theory of output and of Keynes's concept of the marginal efficiency of capital – which was itself derived from the neoclassical schedule. Moreover, the fact that the neoclassical theory of output is synonymous with the neoclassical theory of value means that an effective critique of the latter necessarily constitutes an effective critique of the former. There is no logically consistent foundation to the idea that variation in relative prices, or in the rate of interest, or in money wages, will cause the system to tend to a full-employment level of output. Keynes's utilisation of the notion of a demand schedule for investment may perhaps be explained by the pioneering nature of the *General Theory*, in which the main propositions of a new theory of output are combined with vestiges of the old theory; by the need to present an apparently 'complete' theory; and by the pragmatic ambiguity with which many neoclassical propositions were presented in the then dominant Marshallian formulation.

However, once the corrosive influence of the presence of a marginal efficiency of capital schedule is removed, not only is the neoclassical synthesis seen to be without logical foundation (as in any other version of pseudo-Keynesian theory, such as that of Leijonhufvud (1968) or Malinvaud (1977), which assumes a monotonic inverse relationship between the rate of interest and the volume of investment), but also Keynes's positive contribution, the principle of effective demand, is thrown into more dramatic relief.

See Also

- Internal Rate of Return
- Investment and Accumulation
- Investment Decision Criteria
- Keynes, John Maynard (1883–1946)
- Pay-Off Period

Bibliography

Eatwell, J. 1976. Irving Fisher's 'rate of return over cost' and the rate of profit in a capitalist economy. In *Essays in modern capital theory*, ed. M. Brown, K. Sato, and P. Zarembka. New York/Oxford: Elsevier/North-Holland.

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

Garegnani, P. 1978. Notes on consumption, investment and effective demand: I. *Cambridge Journal of Economics* 2(4): 335–353.

Garegnani, P. 1979. Notes on consumption, investment and effective demand: II. *Cambridge Journal of Economics* 3(1): 63–82.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.
Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. New York: Oxford University Press.
Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Blackwell.
Samuelson, P.A. 1962. Parable and realism in capital theory: the surrogate production function. *Review of Economic Studies* 29: 193–206.
Solow, R. 1963. *Capital theory and the rate of return*. Amsterdam: North-Holland.
Solow, R. 1967. The interest rate and the transition between techniques. In *Socialism, capitalism and economic growth*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Marginal Productivity Theory

R. Dorfman

JEL Classifications
D2

Marginal productivity theory is an approach to explaining the rewards received by the various factors or resources that cooperate in production. Broadly stated, it holds that the wage or other payment for the services of a unit of a factor is equal to the decrease in the value of commodities produced that would result if any unit of that factor were withdrawn from the productive process, the amounts of all other factors remaining the same.

The basic justification of this assertion is highly intuitive. It rests on three assumptions: that the product is sold and the factor services are purchased in competitive markets; that the firms in those markets operate so as to maximize their profits; and that the products sold are produced by technologies that satisfy the 'law of variable proportions', which holds that successive equal increments of one factor of production, the amounts of all other factors remaining unchanged, will yield successively smaller increments of physical output. It follows immediately from these assumptions that if the wage of any factor exceeds the value of the output that would be lost if a unit less of that factor were employed, then a unit less of that factor will be employed, and successive units will be released until the inequality is annihilated. Similarly, if the wage of any factor is less than the value of the output that an additional unit could produce, successive units of that factor will be employed until the inequality vanishes.

The motivating concept in the foregoing argument was the effect on the value of output of small changes in the quantities used of different factors of production. This idea is so important that a special vocabulary has developed in order to discuss it with precision. The marginal product of a factor of production is the ratio of the greatest change in the output of some product that can be obtained by a small change in the use of the factor to the change in the use of the factor. The marginal product multiplied by the price of the product is the value of the marginal product. Marginal productivity theory holds that the payment for any factor of production tends to be about equal to the value of its marginal product, where, in a multiproduct firm, the product used in the calculation is the one for which the value of marginal product is greatest.

Clearly, the marginal product and its value may depend on the size of the 'small' change in the amount of the factor that is used in the calculation. To avoid being ambiguous when the amount of the factor is a continuous variable, the concept of marginal productivity is used: the marginal productivity of a factor is the limit that its marginal product approaches as the change in the quantity of the factor approaches zero. The result of multiplying the marginal productivity by the price of the product is called, somewhat inaccurately, the value of the marginal product; confusion rarely results.

Two of the assumptions made above to justify the marginal productivity doctrine can be relaxed. First, if the assumption that the firm produces for a competitive market is dropped, the conclusions of

the theorem has to be weakened slightly. If a firm produces for a market that is not perfectly competitive, it will recognize that it cannot change the quantity of any of the commodities it sells without simultaneously changing the price. Consequently, it will take account of the fact that if it changes the amount used of any factor, the resulting change in sales revenue will not equal the value of the factor's marginal product, but that value adjusted for the induced change in price. The ratio of the change in sales revenue to the change in the employment of a factor, for 'small' changes, is called the factor's marginal revenue product. Then the reasoning used to deduce the marginal productivity doctrine leads to the conclusion that the firm will employ each factor at the level where its marginal revenue product equals its rate of pay, whether or not the firm sells in a competitive market. In competitive markets, the marginal revenue product of a factor equals the value of its marginal product, but not necessarily in other market types.

The assumption that firms operate so as to maximize their profits can also be weakened for some purposes. If the firm operates only so as to produce its outputs at the lowest possible total cost, the same line of argument shows that the rates of pay for any two factors used by the firm will be proportional to the marginal revenue products of the factors. This is a weaker conclusion than was found for profit- maximizing firms, and does not imply any particular relationship between a factor's rate of pay and its marginal revenue product or the value of its marginal product.

Development of the Concept

Simple as it may appear, the marginal productivity principle was seen clearly only after a long, slow evolution. It was first presented in essentially its modern form around 1890, by J.B. Clark and Alfred Marshall, who apparently arrived at it independently. Their formulation built on the work of numerous predecessors, each of whom saw an important aspect of the principle but did not perceive its full generality.

The problem that gave rise to the marginal productivity principle – to explain the distribution of the national income among the great social classes and, especially to explain the shares claimed by the owners of capital and land – was at the top of the agenda of 19th-century economics. Thus, originally, only three very broad factors of production were considered: land, labour and capital, corresponding to the three social classes.

The first application of the principle occurred in the Malthus–Ricardo theory of rent, in particular in the concept of the intensive margin, which held that doses of labour and capital (in unspecified proportions) would be applied to each parcel of land until the value of the increase in product equalled the cost of the dose. The separate rewards to labour and capital were explained on other grounds.

In 1833, Longfield argued that the rate of interest was governed by the earnings of the least productive unit of capital, using a marginal argument. But he did not extend the reasoning to wages. At around the same time, von Thünen applied the principle to both wages and interest but did not publish his findings until much later, and then so obscurely that they had no influence. Jevons, in 1871, accounted for the rate of interest by a marginal argument, but explained wages as a residual after rent and interest were paid. Indeed, Jevons's theory is remarkably similar to Longfield's.

The ingredient that all these applications of the marginal principle missed was that the equality of marginal product and factor reward applied to all factors. Walras in 1874 (and, indeed, J.-B. Say three-quarters of a century before) insisted on treating the various factors of production symmetrically, but he did not derive any of the factor shares by a marginal argument until the later editions of the *Elements*, and then only awkwardly. Thus the marginal insight was not applied symmetrically to all factors until Clark published the papers that led to his *The Distribution of Wealth* (1899), and Marshall published his *Principles of Economics* (1890), thereby introducing a unified theory of income distribution.

The achievement of the unified theory raised a puzzling question: if each unit of every factor was

paid the value of that factor's marginal product, would the total value produced be neither more nor less than just sufficient to make all the factor payments? In 1894 Philip Wicksteed showed that the answer is affirmative for production processes with constant returns to scale, thus establishing the internal consistency of the marginal productivity principle. (Clark had believed it all along, but on inadequate grounds.) Wicksteed's proof amounted to an independent rediscovery of part of Euler's Theorem for Homogeneous Functions.

Beginning with the late 1880s, when the various partial glimpses of the doctrine congealed, marginal productivity theory became an essential part of the accepted explanation of the general level of wages and of the rate of interest, with important implications for practical economic issues. For example, it is often held that unions are powerless to raise the average level of wages because wages are governed by the marginal productivity of the labour force, which union activity cannot affect.

Although the marginal productivity concept was originally applied to explain the rewards of the broad social classes – the workers, landowners and entrepreneur–capitalists – beginning with Walras it became absorbed into the general theory of production and value. In that context it is used to explain the payments for the services of all the classes of factors that enter into production, and the definitions of these classes can be chosen freely to fit the problem under study. The tripartite classification continues to be used frequently, however.

Qualifications

The marginal productivity doctrine does not purport to be a complete explanation, even in principle, of the payments received by factors of production. As the simple, basic argument indicated, it explains only the amount of each factor that an enterprise will employ at different rates of payment for its services and in the presence of given quantities of the other factors used; that is, the demand curves for the factors. Supply curves also are needed to complete the explanation of the equilibrium level of use and rate of payment for the factors.

Furthermore, especially in the version that deals with numerous factors, rather than just two, a high degree of simultaneity arises. The demand curve for each factor depends on the amounts used of the other factors, but those amounts, depend on the amount used of the first factor, so, in the end, the rates of payment and the quantities used of all the factors are determined simultaneously. Consequently, the rates of payment for the various factors cannot be explained except in the context of a full-fledged general equilibrium model. Still, in such a model it often turns out that the payment received by each factor corresponds to its marginal productivity in each productive process in which it is used and in which marginal productivity is a well- defined concept. These complications will be clarified by considering a more formal and rigorous derivation and statement of the principle.

Formal Derivation of the Marginal Productivity Thesis

The theory is based on the behaviour of a profit-maximizing firm in a competitive industry. To describe that behaviour, imagine a firm that produces m products by the use of n factors or inputs. Suppose that the price of the ith product is p_i and that the quantity produced (per year) is y_i. Then the gross revenues per year will be $R = \sum p_i y_i$. Similarly, let w_j be the price per unit of the jth factor used. If the jth factor is a kind of labour or a purchased input w_j is simply its price or wage, but if it is a kind of fixed capital, then w_j should be regarded as its rental cost, normally interest on its purchase or construction cost plus a depreciation allowance. The amount of the jth factor used will be denoted by x_j. Then the total cost incurred per year will be $C = \sum w_j x_j$. The profit that the firm seeks to maximize is $R - C$.

The quantities (per year) of output, y_i and input x_j, cannot be chosen freely. This basic presentation will be limited to the simplest situation, in which the choices are constrained only by an explicit, differentiable production function, which will be

written $g(y1, \ldots, ym, x1, \ldots, xn) = c$. The implicit constraint that none of the arguments of $g(.,.,.. .)$ can be negative has important consequences that will be noted below.

In this set-up, invoking the assumptions mentioned in the second paragraph of this entry, the necessary conditions for a choice of y and x to maximize the firm's profits are the familiar marginal equalities. Specifically:

1. Marginal rates of substitution. The marginal rate of substitution between two factors, say the jth and the kth is the rate at which small amounts of the jth factor can be substituted for the kth with no effect on the rates of output in accordance with the production function constraint. Denote it by $\delta xj = \delta xk$: Mathematically, it is the ratio of the partial derivatives of the production function, or $\delta x_j = \delta x_k = -(\partial g/\partial)/(\partial g/\partial x_k)$ Economically, when the firm's profits are being maximized $\delta x_j/\delta x_k = w_k/w_j$. The intuitive content is clearest when the maximizing condition is written as $w_j\delta x_j = w_k\delta x_k$, which requires that when profits are being maximized the amounts of factors that can be substituted for each other in accordance with the production function must have equal monetary value.
2. Marginal rates of transformation. There is a similar relationship among the rates at which the outputs can be 'transformed' into one another in accordance with the production function constraint. Let $\delta y_i/\delta y_k$ denote the ratio at which production of the ith output can replace production of the kth. Mathematically, $\delta y_i/\delta y_k = (\partial g/\partial y_k)/(\partial g \partial y_i)$. Economically, when profits are being maximized $\delta y_i = \delta y_k = p_k/p_i$. Again, this result asserts that when profits are being maximized a small quantity of one of the outputs can be replaced by a quantity of equal value of any of the other outputs.
3. Marginal productivity of a factor. The final necessary marginal equality relates small changes in the quantity of an input, say x_j, to the resulting changes in the quantity of any of the outputs, say yi, in accordance with the production function. Mathematically, $\delta y_i/\delta x_j = (\partial g/\partial x_j/\partial g/\partial y_i)$. When profits are being maximized $\delta y_i/\delta x_j = w_j/p_i$. Economically, this is seen to require that if any output is increased by a small amount, the use of some factor must be increased by an amount of equal value.

This third differential equality, of course, is the marginal productivity doctrine, which is seen to be one of the consequences of the theory of profit maximization under competitive conditions. It is often written in the form $\text{VMP}_j = p_i(\delta y_i/\delta x_j) = w_j$ for all values of i. This formula defines the value of the marginal product of the jth factor to be the increase in the value produced of any product for which that factor is used, per unit increase in the use of the factor, and holds that the price per unit of that factor's services will be equal to the VMP when profits are being maximized.

Evaluation

At present the marginal productivity principle is used to explain the demand for factors of production in both a two-factor version using aggregate capital and aggregate labour as the factors, and an n-factor version, where n is the number of distinguishable factors used in the production process.

To use the two-factor version it is necessary to establish quantitative measures of the aggregates of dissimilar objects that are given the names 'capital' and 'labour', a task that has never been performed to anyone's satisfaction. For a long time, until the publication of J. Robinson's disturbing paper, 'The production function and the theory of capital' (1953), the lack of satisfactory measures of the aggregates was regarded as a technicality that did not affect the essential insight. But that paper drove home the realization that in the absence of those measures the marginal productivities, i.e. $\partial g/\partial K$ and $\partial g/\partial L$, were essentially undefined. From that time forth, analyses that use the two-factor version have been regarded as simplified 'parables', useful for making an intuitive point, but not to be taken literally.

Clarifying the meaning of 'capital' and 'labour' regarded as homogeneous factors of production continues to be one of the main problems of capital theory.

The *n*-factor version avoids the impossible task of aggregating apples and bulldozers, but has problems of its own. The formulation considered in this article is too simple to fit most industrial or commercial situations. It presumes that the constraints on choices can be described reasonably well by a single well-behaved, differentiable production function. This is generally not the case. Extreme examples are production functions in which factors are used or outputs are produced in fixed proportions. Any cooking recipe provides an example. More usual are cases in which the choices of input and output quantities are constrained by several functional relationships. The typical example is a firm or industry in which several different machines are each used in the production of several different products. Then there will be a functional relationship for each type of machine, to express the capacity of that type required for each combination of quantities of the different products. This is the sort of problem that has given rise to the use of linear programming in production planning and economic planning generally.

Where there are several constraints, the formulation used above does not apply because, essentially, the amounts of the inputs and outputs cannot be varied two at a time if they are connected by more than a single constraint. Marginal productivity is still a well-defined concept, but it no longer satisfies simple formulas like $\mathrm{MP}j = p_j(\delta y_i/\delta x_j)$ for any value of i. Instead, the marginal productivities, as defined above, are identified with the shadow prices in the solution to a mathematical programming problem, which is a considerably less intuitive concept.

Very frequently, if the problem of finding the combination of factor inputs that maximizes profits is solved straightforwardly, some of the input levels in the solution turn out to be negative – which is nonsense. Then, again, resort must be had to mathematical programming types of formulation and interpretation. The essential perceptions of marginal productivity theory still apply, but they can no longer be expressed by equalities between price ratios and ratios of marginal changes.

See Also

- ▶ Capital Theory (Paradoxes)
- ▶ Clark, John Bates (1847–1938)
- ▶ Classical Distribution Theories

Bibliography

Expositions of marginal productivity theory can be found in any standard text on microeconomics, for example Mansfield (1985). A famous and thoughtful presentation is contained in Hicks (1932). More advanced, and more mathematical, treatments, can be found in Baumol (1977) and Malinvaud (1972). For the relation between marginal productivity and mathematical programming, see Dorfman et al. (1958). The standard, and excellent, reference on the history of the doctrine is Stigler (1941).

Baumol, W.J. 1977. *Economic theory and operations analysis*. Englewood Cliffs: Prentice-Hall.

Dorfman, R., P.A. Samuelson, and R.M. Solow. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill.

Hicks, J.R. 1932. *The theory of wages*. New York: Macmillan.

Malinvaud, E. 1972. *Lectures on microeconomic theory*. Amsterdam: North-Holland.

Mansfield, E. 1985. *Microeconomics: Theory and applications*. New York: Norton.

Stigler, G.J. 1941. *Production and distribution theories: The formative period*. New York: Macmillan.

Marginal Revolution

Roger E. Backhouse

Abstract

The marginal revolution saw the introduction of the idea of marginal utility into economics in the early 1870s by Jevons, Walras and Menger. This change in economic theory was a slower process than the word 'revolution' suggests, and, to understand the changes associated with it, it is necessary to explore the scientific, social and political context in which they occurred.

Keywords

Clark, J. B.; Classical economics; Cliffe Leslie, T. E.; Collectivism; Cournot, A. A.; Energetics; Eugenics; Evolution; Evolutionary psychology; Fabian economics; Green, T. H.; Income distribution; Individualism; Ingram, J. K.; Jevons, W. S.; Marginal revolution; Marginal utility; Marshall, A.; Marxism; Mathematics and economics; Menger, C.; Mill, J. S.; Mirowski, P.; Poverty; Rau, K. H.; Say, J.-B.; Schumpeter, J. A.; Simultaneous equations; Social Darwinism; Socialism; Statistics and economics; Steuart, J.; Subjective theory of value; Supply and demand theory of value; Utilitarianism; Walras, L.; Wicksell, J. G. K.

JEL Classifications

B1

The marginal revolution (sometimes called the marginal utility revolution) refers to the introduction into economics, in 1870–1, of the concept of marginal utility by William Stanley Jevons, Léon Walras and Carl Menger and which has widely been seen as involving a revolutionary break with the 'classical' economics of David Ricardo, John Stuart Mill and many of their contemporaries (see Blaug 1996, ch. 8). The value of a commodity was no longer explained in terms of its cost of production (possibly reducible to the labour required to produce it) but in terms of its value to the consumer. The concept of utility was used to explain consumer choices, marginal utility being seen by some (though not all) authors as replacing cost of production as the foundation on which the theory of value rested. In the 1890s marginal techniques were then applied systematically to the problem of income distribution. This change, it is argued, revolutionized economics, laying the foundations on which modern economic theory is built. Its many dimensions – viewing behaviour as optimization, using utility to describe individual behaviour, focusing on individual agents, the use of mathematics – attest to its importance (Hutchison 1978, provides a longer list; Maas 2005, ch. 1, sketches more recent attempts to choose between them). There is disagreement over the extent to which the change should be described as a revolutionary or as an evolutionary change going back many decades, and over its exact significance; but the marginal revolution is firmly established in histories of economic thought. However, while it describes certain developments in economic theory, to understand the changes that took place in economics around this time one should place it in a broader historical context.

Varieties of Marginalism

The most important qualification to the idea of a marginal revolution is the heterogeneity of economics during this period. Classical ideas were dominant in Britain, but even within classical economics there was great variety, and it has even been argued that marginalist ideas can be traced back as far as Steuart (see ▶ English School of Political Economy). At some time, virtually every element in the classical system outlined above had been challenged, many of these challenges leaving their mark. Much of this variety was captured within Mill's *Principles of Political Economy,* which went through seven editions between 1848 and 1873, and was undoubtedly the leading treatment of the subject: he worked with a very broad supply and demand theory of value and had accommodated many modifications to the Ricardian theory of income distribution. From Mill, the jump to marginalist theories was much easier than from Ricardo. Indeed, Alfred Marshall, though unfairly praising Ricardo at the expense of Mill (see O'Brien 1990), derived his theory of value by translating Mill into mathematics during the 1860s; when he encountered Jevons's work, it was a simple matter to graft marginal utility on to a mathematical treatment of supply and demand (Whitaker 1975).

There was also great variety across countries. In Ireland, it has been argued that an independent tradition of subjective value theory had been established at Trinity College Dublin, by successive holders of the Whately Chair (Black 1945). Ireland also produced two leading exponents of a

historical approach to economics, T.E. Cliffe Leslie and John Kells Ingram. Leslie's assault on deductive theorizing was a significant factor in the shaking of confidence in classical economics in Britain in the 1870s (see Hutchison 1953). In Germany, supply and demand theories had a long history, a supply and demand diagram having been used in a textbook as early as 1843 by Heinrich Rau (see Streissler 1990). In France, Smithian political economy had been mediated not through Ricardo but through Jean-Baptiste Say. The work of Augustin Cournot and the engineers of the École des ponts et chaussées, whose analysis rested on the concept of a demand curve, created an intellectual background very different from that prevailing in Britain.

These differences, together with profound differences in their personal backgrounds, meant that the work of Jevons, Walras and Menger, though often bracketed together, was far from homogeneous (see Jaffe 1975). Though Jevons and Walras both advocated the importance of mathematical argument, their emphases differed. Walras, closer to French rationalism, saw his general equilibrium equations as an abstract system that could solve the same problem that was solved in the real world by other means. Jevons focused on mechanical analogies and the notion that the same methods could be applied to physical and social sciences (Maas 2005). The contrast was even more marked in their applied work, where Jevons was a pioneer in the use of statistics but Walras was not. Menger, in contrast with both of them, rejected the use of mathematics, seeing the use of simultaneous equations as incompatible with identifying the causal relations between human needs and the value of commodities.

The varieties of marginalism increase further when later marginalists are brought into the account. Jevons, Walras and Menger did have disciples, most of them took their analysis in new directions and many are in many ways originals, the best examples being Marshall, Joseph Alois Schumpeter (Austrian, geographically and intellectually, yet an admirer of Walras), Knut Wicksell (whose Swedish synthesis of Austrian and Walrasian ideas bore little resemblance to Schumpeter's) and John Bates Clark (who constructed a non-mathematical American version of marginalism). Given this variety, it is not surprising that the marginal revolution can also be portrayed as a very slow process. Even in the 1880s and 1890s, some economists were still writing textbooks organized on classical lines, marginalist ways of thinking co-existing with other lines of enquiry.

The Wider Context

While scholars might, at one time, have been content to explain the advent of marginalist ideas in terms of economists coming to see the truth about consumers and value, historians are no longer satisfied with such explanations, arguing that economic ideas have to be explained in terms of the context out of which they arose. One context is that of 19th-century science. The most widely discussed explanation has been Mirowski's (1984, 1989) argument that marginalist economics reflects developments in physics (see De Marchi 1993). The 1860s saw the rise of energetics – the attempt to reduce all physical phenomena to energy. If physical phenomena could be reduced to energy, then so should social phenomena. More than that, adopting the methods of physical scientists and the mathematics of maximization and energy conservation offered economists the possibility of acquiring the status of physicists, adopting similar standards of rigour. Mirowski directed historians' attention to the many passages where Jevons, Walras and others stated explicitly that this was what they were doing. Though Mirowski drew normative conclusions about which many historians have been sceptical, and though his interpretation clearly does not fit some of the most important marginalists (notably Menger and Marshall), historians have taken up the idea that a major dimension to the marginal revolution was seeing economics as amenable to the methods of the physical sciences rather than as something radically different (see Maas 2005; Schabas 2005).

Moreover, at this period, physics was not the only prestigious natural science: controversies over evolution were at their height, following the publication of Charles Darwin's *Origin of Species* and

the application of evolutionary arguments to human society by Herbert Spencer. This cannot explain the advent of marginalism, but it represents an important additional connection between economics and contemporary science and helps explain why economics looked very different at the end of the 19th century from the way it looked in the 1860s. Raffaelli (2003) has pointed out that Marshall, perhaps the most significant figure in late-19th century marginalist economics, based his economics, not on the Benthamite utilitarianism used by Jevons, but on evolutionary psychology. Human nature was moulded by experience. Evolutionary ideas thus reinforced the notion that human beings had to be seen as different from one another and that they could be changed. This way of thinking could lead into eugenics, a widely entertained body of ideas that developed towards the end of the century (see Peart and Levy 2005). But such ideas also served to undermine the Malthusian bogey that had provided an argument against much social reform throughout the century (Stedman Jones 1984). Marshall, for example, though he used the static, mechanical apparatus of supply and demand, used it to discuss dynamic processes. He saw industries evolving as biological species, and human character changing in response to human activities, a process that was too complicated to be represented mathematically, and as a result never worked with formal dynamic models: they would have been too mechanical. Against such arguments that evolution became influential at that time, Schabas (2005), though stressing that neoclassical economists were very interested in psychology, has recently questioned whether Darwin has as much influence as has been claimed.

The significance of evolutionary ideas points to another aspect of the context against which the advent of marginalist ideas needs to be set: the political climate. The late 19th century has been called the liberal age, when Europe moved towards freer trade and the franchise became more democratic. The progress of liberal ideas and policies varied greatly from country to country, but everywhere there was debate over the merits of liberalism and collectivism, with the latter taking many forms, ranging from Fabian 'municipal socialism' to Marxian socialism. In Britain, the mid-century radicals, amongst whom Mill was pre-eminent, were liberals who wanted to reform the institutions of society in ways consistent with their liberalism. But, by the end of the century, following the extension of the franchise to much of the working class in 1867 and 1884, radicalism became increasingly collectivist. Against Social Darwinist arguments for individualism were ranged ethical arguments for reform, from the American Social Gospel movement to the variety of movements for social reform inspired in Britain by the Oxford philosopher T.H. Green (see Richter 1964). In the same way that the Great Depression motivated many who came into economics in the 1930s and 1940s, the problem of poverty affected this earlier generation. Economists' attitudes towards policy changed (see Hutchison 1978), as did the way they developed their theories, the most noticeable example being the development of welfare economics by the Cambridge School, J.A. Hobson, and others.

Though it was again a process the speed of which varied greatly from country to country, a further element of the context in which the marginal revolution took place was the professionalization of economics. By the middle of the 19th century, economics was being developed by a mixture of academics and members of a broader educated elite; those recognized as economists might be politicians or businessmen. Specialist journals existed in some countries, but original work in the subject was also published in journals read by non-specialists. By the end of the century economics, like many other disciplines, had changed, becoming an academic discipline in which the main communication was between specialists. This made possible a different type of discourse, more technical and addressing issues that might seem more tenuously related to issues of concern to lay people.

Conclusions

The marginal revolution, like other revolutions in economics, is associated with changes in economic theory that undoubtedly altered the way economics was conceived. However, picking out

a single theoretical or methodological innovation that explains why the marginal revolution was apparently so important has proved difficult. The reason may be that, as in the case of the Keynesian Revolution, though economics changed profoundly in the closing decades of the 19th century, these changes owed as much, if not more, to deeper changes in the social, political and intellectual context in which economists were working as to any specific innovation in economic theory.

See Also

- ▶ English School of Political Economy
- ▶ Jevons, William Stanley (1835–1882)
- ▶ Marshall, Alfred (1842–1924)
- ▶ Menger, Carl (1840–1921)
- ▶ Walras, Léon (1834–1910)

Bibliography

Black, R.D.C. 1945. Trinity College Dublin and the theory of value. *Economica* 47: 140–148.

Blaug, M. 1996. *Economic theory in retrospect*, 5th ed. Cambridge: Cambridge University Press.

De Marchi, N. (ed.) 1993. *Non-natural social science: Reflecting on the enterprise of more heat than light*. Durham: Duke University Press. Also in *History of Political Economy* 25(Suppl): 271–282.

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.

Jaffe, W. 1975. Menger, Jevons and Walras dehomogenized. *Economic Inquiry* 14: 511–524.

Maas, H.B.J.B. 2005. *William Stanley Jevons and the making of modern economics*. Cambridge: Cambridge University Press.

Mirowski, P. 1984. Physics and the marginalist revolution. *Cambridge Journal of Economics* 8: 361–379.

Mirowski, P. 1989. *More heat than light: Economics as social physics, physics as nature's economics*. Cambridge: Cambridge University Press.

O'Brien, D.P. 1990. Marshall's work in relation to classical economics. In *Centenary essays on Alfred Marshall*, ed. J. Whitaker. Cambridge: Cambridge University Press.

Peart, S., and D. Levy. 2005. *The vanity of the philosopher: From equality to hierarchy in post-classical economics*. Ann Arbor: University of Michigan Press.

Raffaelli, T. 2003. *Marshall's evolutionary economics*. London: Routledge.

Richter, M. 1964. *The politics of conscience: T.H. Green and his age*. London: Weidenfeld & Nicolson.

Schabas, M. 2005. *The natural origins of economics*. Chicago: University of Chicago Press.

Stedman Jones, G. 1984. *Outcast London: A study in the relationship between classes in Victorian society*, 2nd ed. New York: Pantheon Books.

Streissler, E.W. 1990. The influence of German economics on the work of Menger and Marshall. *History of Political Economy* 22(Annual Supplement): 31–68.

Whitaker, J.K. (ed.). 1975. *The early economic writings of Alfred Marshall, 1867–1890*. London: Macmillan.

Marginal Utility of Money

Eugene Silberberg

Abstract

Alfred Marshall identified the area to the left of a demand curve as consumer surplus, but he added that his discussion was valid only under the assumption of 'constant marginal utility of money'. For much of the 20th century economists debated the meaning of that phrase and its relevance to consumer surplus. The analysis became clear only after the development of duality theory, particularly the properties of the expenditure function. Marshall's caution becomes unnecessary with a proper definition of consumer surplus.

Keywords

Consumer surplus; Envelope theorem; Hicksian and Marshallian demands; Homothetic utility functions; Indirect utility function; Marginal utility of money; Marshall, A.; Money; Roy's equality

JEL Classifications

D11

Interest in the marginal utility of money probably dates from Alfred Marshall's identification of consumer surplus as the area under the demand curve. Marshall went on to add a qualification to his analysis:

> In the same way if we were to neglect for the moment the fact that the same sum of money represents a different amount of pleasure to different people, we might measure the surplus satisfaction which the sale of tea affords, say, in the London market, by the aggregate of the sums by which the prices shown in a complete list of demand prices of tea exceeds its selling price. (Marshall 1920, p. 106)

In the mathematical appendix (Note VI), Marshall identifies the 'total utility of the commodity' with the area under the demand curve, defined by an integral, but then qualifies that analysis by saying 'we assume that the marginal utility of money to the individual purchaser is the same throughout'. The meaning of these phrases is anything but clear. The text phrase seems to indicate that interpersonal comparisons of utility are a necessary prerequisite for the use of consumer's surplus; in the appendix, Marshall's concern is that, as more of a commodity is purchased, money will yield less satisfaction to the consumer, destroying any linear relationship between money and utility.

Later interpretation of 'constant marginal utility of money' was further complicated by the use of the word 'money' in two different contexts. To Marshall, money provided no direct utility to the consumer; it was a device solely for lowering the transactions cost of exchange. The concurrently developed general equilibrium theory of Walras, however, treated money as that one good which happened to have the additional property of serving as the medium of exchange, a numéraire commodity whose price was unity.

We now analyse the connection between the marginal utility of money and consumer's surplus. The consumer is assumed to maximize $U = U(x_1, \ldots, x_n)$ subject to $\sum p_i x_i = M$. We derive the Marshallian (money income held constant) demand functions $x_i = x_i^M(p_1, \ldots, p_n, M)$ along with $\lambda^M(p_1, \ldots, p_n, M)$ using the Lagrangian $L = U + \lambda(M - \sum p_i x_i)$.

The indirect utility function $U^*(p_1, \ldots, p_n, M) = U(x_1^M, \ldots, x_n^M)$ indicates the maximum utility for given prices and money income. Using the envelope theorem, $\partial U^*/\partial M = \lambda^M$, the marginal utility of money income. Also, $\partial U^*/\partial p_i = -\lambda^M x_i^M$ (Roy's identity). The Hicksian (utility held constant) or 'compensated' demand functions $x_i^U(p_1, \ldots, p_n, U)$ are derived from minimizing $M = \sum p_i x_i$ subject to $U(x_1, \ldots, x_n) = U_0$. The expenditure function $M^*(p_1, \ldots, p_n, U^0) = \sum p_i x_i^u$ indicates the minimum cost of maintaining utility level U^0 for arbitrary prices $p_1, \ldots, p_n$ the envelope theorem, the Hicksian demands are the first partials of the expenditure function: $x_i^U = \partial M^*/\partial p_i$. (See ▶ Hicksian and Marshallian Demands.)

The area to the left of a consumer's demand curve between two prices (where the initial price is higher than the final price), is $-\int x_i \partial p_i$. The units of this integral are that of money income, being price times quantity. Suppose this area equals some value A. The issue is: what question does A answer, and what is the relation between that answer and the marginal utility of money? Since a Hicksian demand function is the first partial of the expenditure function, the area to the left of this demand curve is simply a change in the expenditure function:

$$-\int x_i^U \mathrm{d}p_i = -\int (\partial M^*/\partial p_i)\mathrm{d}p_i = M^*(p^0, U^0) - M^*(p^1, U^0)$$

where p^0 and p^1 are the initial and final price vectors over which the integral is taken. The area to the left of Hicksian demand function therefore represents a change in expenditure with utility held constant; this area indicates the amount a consumer would be willing to pay (or have to be paid) to willingly accept some change in the purchase price of some good. Moreover, there is no need to invoke any assumption at all about the marginal utility of money.

The area to the left of the Marshallian demand function, however, has no such easy interpretation, because unlike the Hicksian demands, the Marshallian demand functions are *not* in general the partial derivatives of some integral function; therefore the integrals of the Marshallian demands are not expressible in terms of changes in some well-defined function of the initial and final prices and income levels. From Roy's equality, the Marshallian demands are the first partials of the indirect utility function *divided by the marginal utility of income*. Thus

$$-\int x_i^M \mathrm{d}p_i = -\int \left(1/\lambda^M\right)\left(\lambda^M x_i^M\right)\mathrm{d}p_i = \int \left(1/\lambda^M\right)\left(\partial U^*/\partial p_i\right)\mathrm{d}p_i.$$

However, if the marginal utility of money term is 'constant', that is, independent of prices, it can be moved in front of the integral sign; only then can this expression be integrated to yield a function of the endpoint prices (and money income):

$$-\int x_i^M \mathrm{d}p_i = \left(1/\lambda^M\right)\int \left(\partial U^*/\partial p_i\right)\mathrm{d}p_i = \left(1/\lambda^M\right)\left[U\left(p^1\right) - U\left(p^0\right)\right].$$

Thus, in this case, the area to the left of the Marshallian demand function would equal a change in utility divided by the marginal utility of money, thus converting that change in utility into units of money. Marshall's claim that the area to the left of a demand curve may be interpreted as a change in utility under the assumption of constant marginal utility of money would thus be technically correct for the demand functions derived from utility maximization, though how much of the above discussion he had in mind can easily be debated.

The problem with this analysis is that λ^M cannot literally be a 'constant,' as shown by Samuelson (1942). Since $\lambda^M = U_i/P_i$ a proportionate change (for example, doubling) of prices and income leaves the amount of the goods consumed unchanged (since the Marshallian demand functions are homogeneous of degree zero in prices and income), and thus the numerator of this expression unchanged. However, the denominator has doubled, meaning λ^M has halved. Thus $\lambda^M = (p_1, \ldots, p_n, M)$ must be homogeneous of degree –1; it can be independent of at most n of its arguments. It can, for example, be independent of all prices, but not income also, or it can be independent of $n - 1$ prices and income.

Since $\partial U^*/\partial p_i = -\lambda^M x_i^M$ and $\partial U^*/\partial M = \lambda^M$, Young's theorem on invariance of partial derivatives to the order of differentiation yields (omitting superscripts)

$$M^*_{p_i M} = -[\lambda \partial x_i/\partial M + x_i \partial\lambda/\partial M] = \partial\lambda/\partial p_i = M^*_{M p_i}.$$

Suppose

$$\partial\lambda^M/\partial p_i = 0, i = 1, \ldots, n.$$

Then

$$(M/x_i)(\partial x_i/\partial M) = -(M/\lambda)(\partial\lambda/\partial M) \text{ for } i = 1, \ldots, n.$$

That is, the income elasticities are all equal (necessarily to unity, from the budget constraint); thus the utility function must be homothetic. Denoting the Marshallian area CS, we have

$$CS = \left(1/\lambda^M\right)\left[U^*\left(p^1, M\right) - U^*\left(p^0, M\right)\right].$$

Thus for homothetic utility functions, where the indifference curves are all radial blow-ups of each other, the Marshallian area represents the unique monetary equivalent of a change in utility; the coefficient which converts 'utiles' to money income is invariant over the price change.

Suppose now that λ^M is a function of one price only, say p_n. Then from the above equation, $\partial x_i^M / \partial M = 0, i = 1, \ldots, n-1$. Since there is no income effect for goods 1 to $n - 1$, the Marshallian demand functions for those goods coincide with the Hicksian demands. This is the famous case of 'vertically parallel' indifference curves. Therefore the interpretation of the area to the left of any of these Marshallian demand curves is identical to the case of the Hicksian demands, that is, the willingness to pay to face the lower price.

See Also

- ▶ Consumer Surplus
- ▶ Giffen's Paradox
- ▶ Hicksian and Marshallian Demands
- ▶ Indirect Utility Function
- ▶ Marshall, Alfred (1842–1924)

Bibliography

Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.

Samuelson, P.A. 1942. Constancy of the marginal utility of money. In *Studies in mathematical economics and econometrics: In memory of Henry Schultz*, ed. O. Lange, F. McIntyre, and T.O. Yntema. Chicago: University of Chicago Press.

Silberberg, E., and W. Suen. 2000. *The structure of economics*. 3rd ed. New York: McGraw-Hill.

Marginalist Economics

Antonietta Campus

Unsystematic ideas about use value and demand and supply as determinants of the exchange value of commodities, which were developed parallel with, and in opposition to, classical theory, found a systematic treatment at the beginning of the 1870s in W.S. Jevons's *Theory of Political Economy*, C. Menger's *Grundsätze der Volkswirtschaftslehre* (both published in 1871), and Walras's *Eléments d'economie politique pure* (published in two parts in 1874 and 1877). It is usual to mark the beginning of marginalist economics with the appearance of these works, in which the long-sought relationship between use value and exchange value was established for the first time. Earlier works on use value (i.e. utility – reinterpreted in subjective terms) had now led after various elaborations to the principle of decreasing marginal utility (see Dmitriev 1902; Stigler 1950).

What is so new in the works of the 1870s, and of such fundamental importance as to be considered that which 'constitutes the very foundation of the whole edifice of economics' (Walras 1900, p. 44) is the condition of proportionality between prices and marginal utilities for each consumer after exchange, i.e. the condition of maximum utility. This condition (which implies the hypothesis of substitution between goods for each consumer when prices vary) gave an analytical basis to downward sloping demand curves for goods, and, with them, to the idea that, *given* the quantities produced, relative prices are exclusively determined by marginal utilities, independently of the costs of production of commodities.

The 'general relations of demand and supply' as determinants of the normal prices and the value of the component parts of the cost of production, i.e. distribution, were advanced by Walras in 1877 in the second part of his *Eléments*. But Walras's work was not, at that time, widely read because of its mathematical difficulty. Jevons's *Theory* and Menger's *Grundsätze* contained no systematic alternative to the undoubtedly confused classical theory of distribution, in the then dominant form found in J.S. Mill's *Principles*.

The lack of a sufficiently well worked out theory of distribution which could be coordinated with the new theory of value, was reflected in the lack of coordination between costs and prices. Yet there remained the apparent fact that in a competitive economy in the long run, prices tend to be equal to costs. When Marshall reviewed Jevons's *Theory* in 1872, he pointed out these deficiencies. However, he did not apparently have a solution in view at the time, for in 1909 he wrote to Cannan:

> There remained great lacunae in my theory till about '85; when on my return to Cambridge, I resolved to try to find out what I really did think about Distribution and I gradually developed . . . the doctrine of substitution between prima facie non-competitive industrial groups, of quasi-rents, etc. (see Pigou 1925, p. 405).

The eagerness to fill this void must have been considerable. With the publication in 1867 of Volume I of Marx's *Capital*, Ricardo's theory of value and distribution had reappeared, not in the conciliatory form of J.S. Mill's *Principles*, but in the dangerous form which had been typical of the theory in the decade following Ricardo's death. According to Böhm-Bawerk, this theory constituted for the Germany of 1884 'the focal point about which attack and defence rally in the war in which the issue is the system under which human society shall be organized' (Böhm-Bawerk 1884, p. 241).

It was in this context that an attack on Marx's theory of value was launched simultaneously in 1884 by Wicksteed in Britain and by Böhm-Bawerk in Austria. Both beginners as economists at this time, they became in a few years two of the

great makers of the marginalist theory of distribution, following the line laid down in the works of Jevons and Menger. They conducted their criticism of Marx's theory of value along essentially similar lines, which clearly reflect the impasse in the marginalist theory of *distribution* at the time. The chosen line of criticism, which perhaps they were obliged to follow, was of an 'esoteric' nature, i.e. that of simply contrasting the utility theory with the labour theory of value. Böhm-Bawerk's critique, for instance, basically maintained that in Marx's theory, no less than in Ricardo's, labour rather than utility could be singled out as the source of value because the analysis was artificially restricted to reproduceable goods alone. When this restriction was removed, and the wider category of 'economic goods' – whether reproduceable or not – was considered, it would be apparent that utility, not labour, is the common source, and determining element, of exchange value.

The most obvious objection to this line of argument is that, even allowing that marginal utility theory could explain the price of *non-reproducible* goods, it was quite unable at this stage to explain the prices of *reproducible* goods – that is, of those goods the price of which is subject to the constraint of cost. After all, it was Böhm-Bawerk's opinion that to understand the connection between price and cost 'is to understand a good half of economics' (Böhm-Bawerk 1889, p. 249). If we leave aside the publication in 1877 of the second part of Walras's *Eléments* (on account of its total lack of impact at the time), the 'good half of economics' (on which the other half ultimately depended) in 1884 had not been developed yet by marginalists: hence the 'esoteric' nature of their critique of Marx's theory of value.

On the other hand, the economists who opposed marginalist theory, because of the crucial role they assigned to the cost of production, did not make their case well. This was because the notion of cost itself, which, vague enough in J.S. Mill's *Principles* (not least because of the eclectic nature of that work), had become, in Cairnes's *Some Leading Principles* (1874), what Whitaker defined as 'an appalling jumble of ideas' (1904, p. 10). And it was to become even worse in the 1880s, following the abandonment – after Walker's attack in 1875 and 1876 – of the Wages Fund theory, and the spread, during the 1880s, of what Cannan defines as 'the produce-less-deduction' theories of distribution (Cannan 1929, pp. 356–8). This explains why the discussion of economic theory took on a form peculiar to this period of interregnum: that of a frontal opposition between cost theory and utility theory.

Marshall's *Principles* (published in 1890) at last provided a widely comprehensible solution to the problem of the cost–price relationship that the marginalists had been seeking for twenty years. Marshall established the relationship by simultaneously determining by the same principle, prices as determined by the principle of decreasing marginal utility, and the value of the component parts of the cost of production as determined by the analogous principle of decreasing marginal productivity (which had been discovered later than marginal utility and had certainly been prompted by it, as Wicksteed clearly indicated: 1894, pp. 7–10).

The solution to the problem of the cost–price relationship within marginalist theory is obviously not a 'reconciliation' between classical and marginal theory within a more complete theoretical paradigm. The idea of a reconciliation is, rather, the version of the facts that Marshall ably put forward and soon caused to prevail, favoured in this by the then state of confusion of traditional cost theory, and by the peculiar context in which discussions were necessarily conducted at that time (that of frontal opposition between cost theory and utility theory).

In fact, what Marshall pointed out in his *Principles* through his 'doctrines of substitution' between goods and methods, was a new unifying principle of simultaneous determination of the prices and the value of the component parts of the costs of production: the principle of supply and demand. In Marshall's own words:

> The 'cost of production principle' and the 'final utility' principle … are component parts of the one all-ruling law of supply and demand' insofar as 'marginal uses and costs do not govern value but are governed together with value by the general relations of demand and supply (Marshall 1890, p. 280; p. 410).

This solution, which had already been presented by Walras should have been particularly congenial to Marshall, considering that, already in his review of Jevons's *Theory* in 1872, he had pointed out as a basic limit of Jevons's work a 'successivistic' approach, so to speak, to value and distribution, rather than one of simultaneous determination of prices, distribution and quantities produced.

It is of particular importance to note Marshall's statement in a letter to J.B. Clark in 1908: 'My whole life has been and will be given to presenting in realistic form as much as I can of my Note xxi' (see Pigou 1925, p. 417). Note xxi of his *Principles* is – except for the treatment of capital – substantially Walras's general equilibrium system, generalized for variable coefficients. The *Principles* – the work of Marshall's entire life – is thus essentially a presentation 'in realistic form' of the general equilibrium system which we find in Note xxi. An essential premise, in this 'realistic form' of presentation, is the demonstration of the analytical bases of the 'general conditions of demand and supply', i.e., the 'doctrines of substitution' (following on from the principles of decreasing marginal utility and productivity on which the general equilibrium system in Note xxi rests).

The illustration of the analytical bases of demand and supply was perhaps the most important element which, by making it comprehensible, quickly brought acceptance to a theory which, as presented in Walras's *Eléments*, was far from comprehensible at the time and therefore not amenable to practical application. The possibility of practical application of the 'general conditions of demand and supply', i.e. of general equilibrium, was pursued by Marshall fundamentally through his peculiar method of 'partial equilibrium' (Robbins 1932, p. xv and fn3) later to become one of the most debated aspects of Marshall's *Principles* (see Sraffa 1925, 1926; Robertson, Shove and Sraffa 1930; Newman 1960). Whatever the demerits of the 'partial equilibrium' method, it was thanks to this 'realistic form' of presentation that marginalist economics gained its first general acceptance through the pervasive ascendancy of Marshall's *Principles* in Britain and, directly or indirectly, the United States, Sweden and a large part of Europe (see Shove 1942, pp. 313–16).

But, as Robbins puts it, Marshall's 'peculiar blend of realistic knowledge and theoretical insight … was not necessarily conducive to clear presentation of abstract theoretic issues' (Robbins 1934, p. 10). In fact, Marshall's often explicit propensity to evade precisely defined economic notions, with the justification that, in concrete realities, everything 'shades into the other by imperceptible gradations' facilitated, together with the domination of Marshall's version of marginalist economics, that blurring of difficulties which beset the theory from its beginnings and which were amply debated in the period in which the first six editions of the *Principles* were published (1890–1910).

These difficulties can be illustrated in the simple terms in which they first appeared. The discovery of the principle of decreasing marginal productivity (at which, on the analogy of the earlier principle of decreasing marginal utility, Edgeworth, Marshall, J.B. Clark, Wicksteed, Wicksell and Walras himself arrived simultaneously, between the end of 1880s and the beginning of the 1890s) suggested a method, long sought by the opponents of classical economics, through which the product of each agent of production 'may be disentangled from the product of cooperating agents and separately identified' (Clark 1899, p. viii). However, this possibility of 'disentanglement' proved problematical when the attempt was made to 'identify' the product of that peculiar agent of production which is capital. And on the notion of capital to which one must have recourse to determine distribution, on the basis of the marginalist principles of supply and demand, the greatest exponents of the marginalist theory of distribution, from Böhm-Bawerk and J.B. Clark to Walras and Marshall, openly declared themselves to be at variance with each other. In Böhm-Bawerk's words: 'It is an almost tragicomic circumstance that the champions of the different definitions of capital charge each other with the same error, the irrelevance of the recommended concept' (Böhm-Bawerk 1889, 3rd edn, 1909, Bk. I, ch. III, fn96).

While the divergences as to the way of dealing with capital obviously involved differences in the determination of the rate of profit, in fact they implied more pervasive difficulties. Böhm-Bawerk himself rightly observed that: 'when divergence is as wide as it is on this point of capital, we are forced to the conclusion that there must be something quite unusual about this specific apple of discord.' And he added that Knies appraised the implications of the controversy over capital 'quite accurately', when he said that 'there is more involved here than in the ordinary case of a conflict over a felicitous versus an awkward definition, or even a right versus a wrong *definition*' (Böhm-Bawerk 1909, p. 31). In fact, on account of the necessary relationship between the rates of profit, wages, and rent (Wicksteed 1894), the disagreement over the treatment of capital and therefore over the determination of the rate of profit implied difficulties for the whole theory of distribution and thus for the determination of costs, and normal prices. It was to this state of things that Ashley must have been referring in his 1907 Presidential address to Section F of the British Association, when he said: 'There is hardly a single point in the whole theory of distribution on which there is as yet any approach to unanimity.' And – assuredly having Marshall's attitude to controversy in mind – he remarked: 'Doubtless all the differences could be construed as differences of emphasis; but this is hardly reassuring, for the emphasis may differ so much as to give totally opposite impressions'. (Ashley 1907, pp. 477–8).

Given the almost total domination that marginalist economics has enjoyed for about a century, it would seem natural to think that

> from these 'clashes of thought' between marginalist theoreticians 'the spark of an ultimate truth had at length been struck' (Sraffa 1926, p. 535).

However, things did not really go this way. What happened was rather that 'Clark's value concept of capital . . . gained a considerable and constantly increasing number of adherents' (Böhm-Bawerk 1909, p. 57). Alfred Marshall was perhaps the most important of its 'adherents'. With this adhesion, 'considerations of capital theory proper . . . simply disappeared from the picture' in the English-speaking world (Robbins 1934, p. xiv), at least until the 1930s.

The unresolved question – lying at the very foundation of marginalist theory – as to that 'something quite unusual' which 'there must be' about capital, has, however, been brought to full light in 1960, with the publication of Sraffa's *Production of Commodities by Means of Commodities*. One of the important things proved in this work is that, when commodities are produced by means of 'capital' as well as labour and land, there will in general be 'reversals in the direction of the movement of relative prices' when distribution varies (Sraffa 1960, p. 38). This implies that there is no reason why the 'laws of substitution' between goods and methods, which lie at the basis of the demand curves for goods and factors, should go in the direction required to define downward sloping demand curves for factors. With this, the marginalist theory of distribution seems to land in an untenable position – even in the version presented by Walras – and, with it, the connected theory of the cost of production and normal prices – unless we disregard produced means of production and consider some 'model of pure exchange' as acceptable for the explanation of value. This would be tantamount to accepting, after a century of theoretical 'refinements', what was patently unacceptable in marginalist economics between 1870 and 1890: the lack of coordination between price and cost.

See Also

► 'Neoclassical'

Bibliography

Ashley, W.J. 1907. The present position of political economy. *Economic Journal*, December.

Böhm-Bawerk, E. von. 1884. *Capital and interest*, Vol. I. 4th edn, 1921. South Holland: Libertarian Press, 1959.

Böhm-Bawerk, E. von. 1889. *Capital and interest*, Vol. II. 4th edn, 1921. South Holland: Libertarian Press, 1959.

Böhm-Bawerk, E. von. 1909. *Capital and interest*, Vol. II. 3rd edn of first half-volume.

Cairnes, J.E. 1874. *Some leading principles of political economy newly expounded*. London: Macmillan.

Cannan, E. 1929. *A review of economic theory*, 1964. London: Cass.

Clark, J.B. 1891. Distribution as determined by a law of rent. *Quarterly Journal of Economics* 5(April): 289–315.
Clark, J.B. 1899. *The distribution of wealth*, 1965. Reprinted, New York: Kelley.
Dmitriev, V.K. 1902. *Economic essays on value, competition and utility*, 1904 edn. Cambridge: Cambridge University Press, 1974.
Edgeworth, F.Y. 1889. Address to Section F of the British Association. (Quoted in Marshall (1890), 848, and Stigler (1941), 132.)
Ingram, J.K. 1878. The present position and prospects of political economy. Paper read to Section F of the British Association, as reprinted in *Essays in economic method*, ed. R.L. Smith. London: Duckworth, 1962.
Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan. Reprinted, New York: Kelley, 1957.
Jevons, W.S. 1876. The future of political economy. Introductory Lecture at the opening of the session 1876–77 at University College, London. Reprinted in W.S. Jevons, *Principles of economics*. London: Macmillan, 1905; reprinted, New York: Kelley, 1965.
Marshall, A. 1872. Mr Jevons' theory of political economy. *The Academy*, 1 April. Reprinted in *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan, 1925.
Marshall, A. 1890. *Principles of economics*. 9th (Variorum) edn. London: Macmillan, 1961.
Marx, K. 1867. *Capital. A critique of political economy*, Vol. I. London: Lawrence & Wishart, 1954.
Marx, K. 1894. *Capital. A critique of political economy*, Vol. III. London: Lawrence & Wishart, 1977.
Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Vienna: Braumüller. Trans. as *Principles of Economics*. Glencoe: Irwin, 1950.
Menger, C. 1883. *Problems of economics and sociology*. Urbana: University of Illinois Press, 1963.
Menger, C. 1884. *Die Irrthümer des Historismus*. Vienna: Hölder.
Newman, P. 1960. The erosion of Marshall's theory of value. *Quarterly Journal of Economics* 74(November): 587–601.
Pigou, A.C. (ed.) 1925. *Memorials of Alfred Marshall*. London: Macmillan. Reprinted, New York: Kelley, 1966.
Robbins, L. 1932. Introduction to Vol. I of P.H. Wicksteed, *The common sense of political economy* (1st edn, 1910). London: Routledge & Kegan Paul, 1946.
Robbins, L. 1934. Introduction to Vol. I of K. Wicksell, *Lectures on political economy* (1st edn, 1901). London: Routledge & Kegan Paul, 1961.
Robertson, D.H. 1930. Contribution to 'Increasing returns and the representative firm. A symposium'. *Economic Journal* 40(March): 79–116.
Shove, G.F. 1930. Contribution to 'Increasing returns and the representative firm. A symposium'. *Economic Journal* 40(March): 79–116.
Shove, G.F. 1942. The place of Marshall's *Principles* in the development of economic theory. *Economic Journal* 52(December): 294–329.
Sraffa, P. 1925. Sulle relazione fra costo e quantità prodotta. *Annali di Economia* 2(1): 277–328.
Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36(December): 535–550.
Sraffa, P. 1930. Contribution to 'Increasing returns and the representative firm. A Symposium'. *Economic Journal* 40(March): 79–116.
Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory*. Cambridge: Cambridge University Press.
Stigler, G.J. 1941. *Production and distribution theories*. New York: Macmillan.
Stigler, G.J. 1950. The development of utility theory. Pts I and II. *Journal of Political Economy* 58, 307–27; 373–96.
Walker, F.A. 1875. Article in *North American Review*, January.
Walker, F.A. 1876. *The wages question*. New York: H. Holt.
Walras, L. 1874–7. *Eléments d'économie politique pure*. Trans. by W. Jaffé as *Elements of pure economics*. London: Allen & Unwin, 1954.
Walras, L. 1900. Preface to the 4th edn of the *Eléments*. Reprinted in *Elements of pure economics*, ed. W. Jaffé, London: Allen & Unwin, 1954.
Whitaker, A.C. 1904. *History and criticism of the labor theory of value*. Reprinted, New York: Kelley, 1968.
Wicksell, K. 1893. *Value, capital and rent*, 1954. London: Allen & Unwin.
Wicksell, K. 1901–6. *Lectures on political economy*, 2 vols. London: Routledge & Kegan Paul, 1961.
Wicksteed, P.H. 1884. *Das Kapital*: a criticism. *Today*, October. Reprinted in P.H. Wicksteed. *The common sense of political economy*, Vol. II, 1910; reprinted, London: Routledge & Sons, 1946.
Wicksteed, P.H. 1894. *An essay on the coordination of the laws of distribution*. London: Macmillan. Reprinted as No. 12 of London School of Economics, Reprints of Scarce Tracts, London: London School of Economics, 1932.

Marital Institutions

Scott Drewianka

Abstract

Marital institutions are rules governing marriages and divorces. Most work to date has focused on unilateral and no-fault divorce reforms. Theoretical discussions generally

hinge on the applicability of the Coase theorem. Empirical evidence is mixed, but generally indicates that those reforms played only a modest or temporary role in generating trends in marriage, divorce and fertility. There is more consistent evidence of substantial effects on intrahousehold allocation and other distributional outcomes, especially in conjunction with rules on post-divorce division of property. Several new institutions that have emerged in recent years present promising opportunities for future research.

Keywords

Coase theorem; Civil partnerships; Covenant marriage; Divorce; Domestic violence; Marital institutions; Marriage

JEL Classifications

J12; K36; D13

Marital institutions are laws and customs governing relationships between spouses or domestic partners. Examples include rules defining who may marry, the benefits and responsibilities associated with various family arrangements, the circumstances under which a partnership may end, and the division of property and child custody in that event.

Although researchers have studied many such institutions, the oldest and most developed branch of the literature investigates a wave of divorce reforms enacted during the 1960s and 1970s in most US states and several countries. Two common changes involved dropping the requirement that one spouse be guilty of violating terms of the marital contract ('no-fault divorce') and allowing one spouse to obtain a divorce without the other's consent ('unilateral divorce'). Researchers' interest was largely motivated by contemporaneous changes in family structure. The US divorce rate more than doubled and marriage and fertility rates declined during the reform period, leading some to suspect a causal relationship. However, many economists remained sceptical because divorce rates rose steadily before the reform period, increased during the reform period even in places that did not enact reforms, and fell sharply in subsequent decades.

Theory provides additional reason for doubt. Following the logic of the Coase theorem, Becker (1981) argues that if intrahousehold transfers can be negotiated costlessly, divorce occurs only when a marriage no longer generates a surplus. In that view, reforms that reassign bargaining power between spouses would not affect the incidence of divorce. However, Becker's reasoning would break down if negotiation costs were important or if reforms changed the cost of separating. Reforms could also affect the gains from marriage by altering incentives for match-specific investment (Stevenson 2007; Wickelgren 2009).

Numerous papers have attempted to resolve these claims empirically. The main complication is that reforms may be correlated with existing levels or trends in divorce rates, so failing to account for that correlation would lead to biased estimation. For example, the first known empirical study found that no-fault divorce was more common in states with a greater pre-existing demand for divorce (Broel-Plateris 1961).

Such correlations played a central role in early studies using cross-sectional data. Despite using similar data and methods, Peters (1986, 1992) and Allen (1992) reached opposite conclusions about the effect of divorce reforms on US divorce rates. Their dispute revolved around the appropriateness of specifications with regional dummies and controls for pre-reform divorce rates. Peters argued that those variables were necessary to account for the non-random incidence of divorce reforms, but Allen claimed that they unnecessarily removed variation useful for identifying the reforms' effect.

That debate was eventually resolved using panel data. Friedberg (1998) showed that the estimated effect of unilateral divorce on divorce rates was large when state fixed effects were excluded, but statistically insignificant when they were included. She then advanced the issue by showing that reforms were also negatively correlated with trends in divorce. When she added state-specific time-trends to her specification, her results indicated that unilateral divorce laws encouraged divorce.

However, subsequent work by Wolfers (2006) extended Friedberg's model by allowing the law's effect to vary dynamically. His results confirmed that reforms raised divorce rates, but the effect dissipated after about eight years. According to Rasul (2006), the effect is only temporary because the reform alters the optimal sorting of spouses, and divorce rates rose for couples married before the change but not for those married under the new regime.

A related question is whether the reforms affected incentives to start families in the first place. The evidence on marriage is mixed, with some papers claiming a positive effect and others a negative effect. However, none of the estimates is large, and published papers find little evidence of a statistically significant effect.

In contrast, there is more consistent evidence that unilateral divorce slightly reduces overall fertility, slightly increases marital fertility, and reduces nonmarital fertility more substantially. The latter two effects may reflect individuals' increased willingness to marry in the event of a premarital pregnancy if a potential divorce would be easier to obtain (Gruber 2004; Stevenson 2007; Drewianka 2008). Similarly, Ekert-Jaffe and Grossbard (2008) show that a greater share of births are within-marriage in countries with 'community property' laws, which typically grant a greater share of a divorcing couple's assets to the lower-earning spouse.

In sum, while not entirely consistent with the purely Coasian view, most estimates indicate that divorce reforms had only modest or temporary effects on marriage, fertility and divorce rates.

A second branch of the literature investigates distributional outcomes. Even in a Coasian framework, reforms could still alter outcomes related to spouses' relative bargaining power. Indeed, some marital institutions appear specifically intended to affect bargaining power, such as rules governing post-divorce division of assets. However, a critical difficulty in assessing such effects lies in knowing which spouse benefits from the reform. Even when the change clearly benefits the spouse who is more eager to end the relationship (as in the case of unilateral divorce), researchers often cannot be sure which spouse that is.

This challenge may explain conflicting evidence on domestic violence. Dee (2003) finds that wives are more likely to murder their husbands (but not conversely) under unilateral divorce, especially when rules governing property division tend to favour husbands. His interpretation is that wives respond violently to their heightened risk of economic hardship. In contrast, Stevenson and Wolfers (2006) find that unilateral divorce substantially reduces domestic violence, particularly against women. They argue that it provides both an immediate escape to abused spouses and increased bargaining power that may pre-empt abuse.

Research consistently finds that rules on division of property have increased wives' labour supply, however. Intuitively, those laws tend to favour people with work experience, and thus they particularly encourage market work among married women because that group has a relatively low labour force participation rate (Parkman 1992; Gray 1998).

Although the literature on marital institutions has heretofore emphasized divorce law, a promising topic for future research is new marital institutions. Since the mid-1990s, several states and many European nations have created new legal institutions for domestic partners. These institutions are often available to both heterosexual and homosexual couples, and in some places gays can now marry formally. Another new institution is 'covenant marriage', which is much like standard marriage but with more limited grounds for divorce. Three US states currently offer a covenant marriage option, and it has been considered in many others.

Because these institutions are new, little has been written about them. One interesting empirical finding is that couples choosing covenant marriage would appear to be at high risk of divorce – many are young, heterogamous, and from high-divorce communities (Felkey forthcoming). A theoretical model by Drewianka (2004) indicates that new institutions typically could either encourage or discourage marital commitment by any particular couple, depending on whether the individuals are more likely to enter that institution with their current partner or a

potential future partner. This suggests that theoretical predictions are inherently ambiguous, so empirical evidence will be essential in this context.

See Also

- Bargaining
- Becker, Gary S. (Born 1930)
- Coase Theorem
- Family Economics
- Intrahousehold Welfare
- Marriage and Divorce

Bibliography

Allen, D.W. 1992. Marriage and divorce: Comment. *American Economic Review* 82: 679–685.

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

Broel-Plateris, A. 1961. Marriage disruption and divorce law. PhD diss., University of Chicago.

Dee, T.S. 2003. Until death do you part: The effects of unilateral divorce on spousal homicides. *Economic Inquiry* 41: 163–182.

Drewianka, S. 2004. How will reforms of marital institutions influence marital commitment? A theoretical analysis. *Review of Economics of the Household* 2: 303–323.

Drewianka, S. 2008. Divorce law and family formation. *Journal of Population Economics* 21: 485–503.

Ekert-Jaffe, O., and S. Grossbard. 2008. Does community property discourage unpartnered births? *European Journal of Political Economy* 24: 25–40.

Felkey, A.J. Will you covenant marry me? A preliminary look at a new type of marriage. *Eastern Economic Journal*, forthcoming.

Friedberg, L. 1998. Did unilateral divorce raise divorce rates? Evidence from panel data. *American Economic Review* 88: 608–627.

Gray, J.S. 1998. Divorce-law changes, household bargaining, and married women's labor supply. *American Economic Review* 88: 628–642.

Gruber, J. 2004. Is making divorce easier bad for children? The long run implications of unilateral divorce. *Journal of Labor Economics* 22: 799–833.

Parkman, A. 1992. Unilateral divorce and the labor-force participation rate of married women, revisited. *American Economic Review* 82: 671–678.

Peters, H.E. 1986. Marriage and divorce: Informational constraints and private contracting. *American Economic Review* 76: 437–454.

Peters, H.E. 1992. Marriage and divorce: Reply. *American Economic Review* 82: 686–693.

Rasul, I. 2006. Marriage markets and divorce laws. *Journal of Law, Economics, and Organization* 22: 30–69.

Stevenson, B. 2007. The impact of divorce laws on marriage-specific capital. *Journal of Labor Economics* 25: 75–93.

Stevenson, B., and J. Wolfers. 2006. Bargaining in the shadow of the law: Divorce laws and family distress. *Quarterly Journal of Economics* 121: 267–288.

Wickelgren, A. 2009. Why divorce laws matter: Incentives for noncontractible marital investments under unilateral and consent divorce. *Journal of Law, Economics, and Organization* 25: 80–106.

Wolfers, J. 2006. Did unilateral divorce laws raise divorce rates? A reconciliation and new results. *American Economic Review* 96: 1802–1820.

Market Competition and Selection

Lawrence Blume and David Easley

Abstract

There is a long history in economics of using market selection arguments in defence of rationality hypotheses. According to these arguments, rational investors drive irrational investors out of asset markets and profit maximizing firms drive non-maximizing firms out of goods markets. In this article we present the history of these arguments and discuss the literature that examines whether these arguments for market selection, and its impact on efficiency, are correct.

Keywords

Rationality; Market selection; Kelly rule; Incomplete markets

JEL Classifications

C78; L1; G1; C73

Realized positive profits, not *maximum* profits, are the mark of success and viability. It does not matter through what process of reasoning or motivation such success was achieved. The fact of its accomplishment is sufficient. This is the criterion by which the economic system selects survivors:

> those who realize *positive profits* are the survivors; those who suffer losses disappear. (Alchian 1950, p. 213)

Most economic models make use of extreme rationality hypotheses: firms maximize profits with full knowledge of their technology and prices, and investors are subjective expected utility maximizers whose beliefs are correct. Surely some firms and some investors do not always behave as these models hypothesize, but does this matter for predictions of market outcomes? It could be that the aggregation that takes place in supply and demand results in prices and market quantities that agree with the predictions of models using extreme versions of rationality. It could be that, over time, firms and investors learn to behave as these models predict and so market outcomes converge to those predicted by the models. Finally, it could be that markets select for firms and investors who behave 'as if' they are rational. This last defence of the use of rationality is the essence of the quote from Alchian (1950).

There is a long history in economics of using market selection arguments in defence of rationality hypotheses. The early literature focused on selection for profit maximizing firms. Among its best-known proponents is Friedman (1953, p. 22): 'The process of natural selection thus helps to validate the hypothesis (of profit maximization) or, rather, given natural selection, acceptance of the hypothesis can be based largely on the judgment that it summarizes appropriately the conditions for survival.' Of course, even if the selection reasoning is correct, selection can only work over those types of behaviours that are present in the economy. If no firm maximizes profits, then no profit-maximizing firm can be selected. Alchian was acutely aware of this:

> The pertinent requirement – positive profits through relative efficiency – is weaker than 'maximized profits,' with which, unfortunately, it has been confused. Positive profits accrue to those who are better than their actual competitors, even if the participants are ignorant, intelligent, skilful, etc. The crucial element is one's aggregate position relative to actual competitors, not some hypothetically perfect competitors. As in a race, the award goes to the relatively fastest, even if all the competitors loaf. (Alchian 1950, p. 213)

Enke (1951) argued that, at least in competitive industries, the relatively fastest will in fact be profit maximizers, and so in this case selection will lead to the survival only of profit maximizing firms:

> In the long run, however, if firms are in active competition with one another rather than constituting a number of isolated monopolies, natural selection will tend to permit the survival of only those firms that either through good luck or great skill have managed, almost or completely, to optimize their position and earn the normal profits necessary for survival. In these instances the economist can make aggregate predictions *as if* each and every firm knew how to secure maximum long-run profits. (Enke 1951, p. 567)

Similar market selection arguments have been proposed to justify strong rationality hypotheses on the part of investors. Fama argues that:

> dependency in the noise generating process would tend to produce 'bubbles' in the price series . . . If there are many sophisticated traders in the market, however, they may cause these 'bubbles' to burst before they have a chance to really get underway. (Fama 1965, p. 38)

According to Fama, 'A superior analyst is one whose gains over many periods of time are *consistently* greater than those of the market'. This is at least indirectly an argument for market selection and its affect on the efficiency of prices. Cootner was an early, clear proponent of this argument:

> Given the uncertainty of the real world, the many actual and virtual traders will have many, perhaps equally many, forecasts . . . If any group of traders was consistently better than average in forecasting stock prices, they would accumulate wealth and give their forecasts greater and greater weight. In this process, they would bring the present price closer to the true value. (Cootner 1967, p. 80)

In this article we examine the more recent analyses of whether these arguments for market selection, and its impact on efficiency, are correct. We consider in turn, selection over firms and selection over investors.

Selection over Firms

Alchain, Friedman and Enke argue that a profit dynamic will select for firms that, for whatever

reason, maximize profits. Correspondingly, according to this argument, those that do not act as profit maximizers will be driven out of the market. But how is it that non-maximizers are driven out? The implicit idea is that, in the presence of maximizers, the non-maximizers experience losses that deplete their financial capital, which forces them out of the market. The literature has explored two avenues by which losses of financial capital could have this effect. One is that if the firm's operations are financed from retained earnings, then firms that consistently experience losses would eventually exhaust their retained earnings, causing them to vanish. A second argument is that unsuccessful firms will not be able to raise capital in the financial markets, and may not even be able to retain their initial capital. Thus, so this argument goes, the markets will punish unsuccessful firms, which will eventually vanish.

Winter (1964, 1971) and Nelson and Winter (1982) analyse a retained earnings dynamic. They argue that the retained earnings of profit maximizers will grow fastest, and thus these firms will eventually dominate the market. These authors construct a partial equilibrium model in which the 'as if' hypothesis of profit maximization describes the long-run steady state behaviour of firms. In their analysis, prices are fixed and all firms have access to the same technology. This structure leads to the existence of a uniformly most-fit firm, which is selected for by a retained earnings-based investment dynamic.

The early work on market selection was greatly concerned with the meaning of profit maximization when profits are random. Dutta and Radner (1999) directly take up the question of whether markets select for firms that maximize expected profits. Their answer is 'no': the decision rules that maximize the long probability of survival are not those that maximize expected profits. Dutta and Radner's firms are owned by investors who choose how much of the firm's earnings to reinvest in the firms and how much to withdraw as dividends. An expected profit maximizing firm is one that maximizes the expectation of present discounted value of dividends paid to its owners. This policy results in an upper bound on the retained earnings left in the firm, and from this level of retained earnings any firm can experience a string of losses that results in bankruptcy.

There are two parts to the argument for market selection of profit maximizers. First, there is the issue of whether the market selects for profit maximizers. Second, there is the issue of whether in the long run the economy behaves as if only profit mazimizing firms exist. The Dutta and Radner analysis casts doubt about a positive answer to the first question in stochastic settings. Koopmans (1957) cast doubts about a positive answer to the second question even in a deterministic setting. According to his analysis, appealing to an external dynamic process to defend the profit maximization assumption is not a satisfactory way to proceed. Instead, he believed that the dynamic process itself should be modelled. Nelson and Winter (1982, p. 58) were also aware that the co-evolution of firm behaviour and the economic environment resulting from a complete model of the dynamic process could pose problems for the evolutionary defence of profit maximization. They observed that among the 'less obvious snags for evolutionary arguments that aim to provide a prop for orthodoxy' is 'that the relative profitability ranking of decision rules may not be invariant with respect to market conditions'. They do not, however, go on to provide a general equilibrium analysis of the consequences of replacing static profit maximization with a selection dynamic.

Blume and Easley (2002) showed that Koopman's concern about the market selection dynamic in a general equilibrium setting is correct. They show that although only profit maximizers persist in any steady state of the retained earnings dynamic, the long run of the economy need not be well described by assuming that only profit maximizing firms exist. The difficulty arises because of the endogeneity of prices, which causes the relative profitability of firms to depend on the allocation of capital across the firms. As a result, the retained earnings dynamic need not settle down, and efficient firms can be driven out of the market by inefficient firms.

In addition to raising working capital through retained earnings, firms also enter the capital markets. Whether these markets reinforce the market selection hypothesis, as Friedman argues, or

undermine it, depends on how well these markets function. If markets are complete (without the securities created by non-maximizing firms) and investors are expected utility maximizers with rational expectations, then investors would not allocate capital to non-maximizing firms. Such firms would never produce, and the selection hypothesis would be trivially, and instantly, correct. Alternatively, if some investors have incorrect expectations, then they could invest in non-maximizing firms. The fate of these firms depends on the fate of their investors. So, in this case, the question of selection for profit maximizing firms reduces to the question of selection for investors who act as expected utility maximizers with rational expectations.

Selection over Investors

Friedman et al. argue that asset markets will select for rational investors, and that because of this selection, assets will eventually be priced efficiently. Two interesting approaches have been taken to the selection for rational investors question. First, suppose traders use a variety of portfolio rules. Is it the case that traders whose rules are not rational will lose their money to those who do act as if they are rational? Second, suppose that all traders are subjective expected utility maximizers. Is it the case that markets select for those whose expectations are correct, or most nearly correct?

In order to pose these questions precisely rationality has to be defined (see rationality). The selection literature has asked about selection for a very strong form of rationality – expected utility maximization with correct expectations about the payoffs to assets. This is the interesting question because in economies populated by subjective expected utility maximizers whose beliefs are not tied down by a rational expectations hypothesis we have little to say about asset prices. The mere assumption that investors are subjective expected utility maximizers (in the sense of Savage 1951) places no restrictions on the stochastic process of Arrow security prices (Blume and Easley 2005).

Selection over Rules

Consider an intertemporal general equilibrium economy with a collection of Arrow securities and one physical good available at each date. Suppose traders are characterized by their stochastic processes of endowments of the good and by portfolio and savings rules. A savings rule describes the fraction of wealth the trader saves and invests at each date given any partial history of states. Similarly, a portfolio rule describes the fraction of savings the trader allocates to each Arrow security. The savings and portfolio rules that rational traders could choose form one such class of rules; but other, non-rationally motivated rules are also possible.

Three questions arise about the dynamics of wealth selection in this economy. First, is there any kind of selection at all? Second, is it possible to characterize the rules which win? Third, if selection does take place, does every trader using a rational rule survive, and in the presence of such a trader do all non-rational traders vanish?

In repeated betting, with exogenous odds, the betting rule that maximizes the expected growth rate of wealth is known as the Kelly rule (Kelly 1956). The use of this formula in betting with fixed, but favourable odds was further explored by Breiman (1961). In asset markets the 'odds' are not fixed; instead they are determined by equilibrium asset prices, which in turn depend on traders' portfolio and savings rules. Nonetheless, the market selects over rules according to the expected growth rate of wealth share they induce. Blume and Easley (1992) show that if there is a unique trader using a rule that is globally maximal with respect to this criterion, then this trader eventually controls all the wealth in the economy, and prices are set as if he is the only trader in the economy. A trader whose savings rate is maximal and whose portfolio rule is, in each partial history, the conditional probability of states for tomorrow has a maximal expected growth rate of wealth share. This rule is consistent with the trader having logarithmic utility for consumption, rational expectations and a discount factor that is as large as any trader's savings rate. Thus, if this trader exists, he is selected for. However, rationality alone does not guarantee a maximal expected growth rate of

wealth share. There are rational portfolio rules that do not maximize fitness (even controlling for savings rates), and traders who use these rules can be driven out of the market by traders who use rules that are inconsistent with rationality.

Amir et al. (2005) and Evstigneev et al. (2006) take an alternative approach to selection over rules in asset markets. They consider general one-period assets and ask if there are simple portfolio rules that are selected for, or are evolutionarily stable, when the market is populated by other simple (not explicitly price dependent) portfolio rules. In this research, either all winnings are invested, or equivalently, all investors are assumed to invest an equal fraction of their winnings. So selection operates only over portfolio rules. Amir et al. (2005) find that an investor who apportions his wealth across assets according to their conditional expected relative payoffs drives out all other investors as long as none of the other investors end up holding the market. This result is consistent with Blume and Easley (1992) as the log optimal portfolio rule agrees with the conditional expected relative payoff rule when only these two rules exist in the market. Hence, both these rules hold the market in the limit. Evstigneev et al. (2006) use notions of stability from evolutionary game theory to show that the expected relative payoffs rule is evolutionarily stable.

Selection Among Subjective Expected Utility Maximizers

DeLong et al. (1990, 1991) analyse selection over traders who are subjective expected utility maximizers with differing beliefs. In an overlapping generations model they show (1990) that traders with incorrect beliefs can earn higher expected returns, because they take on extra risk. But as survival is not determined by expected returns, this result does not answer the selection question. DeLong et al. (1991) argue that traders whose beliefs reflect irrational overconfidence can eventually dominate an asset market in which prices are set exogenously. This result appears to contradict Alchian's and Friedman's intuitions. But, as prices are exogenous, these traders are not really trading with each other; if they were, then were traders with incorrect beliefs to dominate the market, prices would reflect their beliefs and rational traders might be able to take advantage of them.

In an economy with complete markets and traders who have a common discount factor, Alchian and Friedman's intuition is correct. Sandroni (2000) shows, in a Lucas trees economy with some rational-expectations traders, that if traders have a common discount factor, then all traders who survive have rational expectations. Blume and Easley (2006) show that this result holds in any Pareto optimal allocation in any bounded classical economy and thus for any complete markets equilibrium. To see why the market selection hypothesis is true for these economies suppose that states are iid and that traders have differing, fixed iid beliefs. Then each trader assigns zero probability to almost all the infinite sample paths that any other trader believes to be possible. Each trader would be willing to give up all his endowment on the sample paths he believes to be impossible in order to obtain more consumption on those he believes to be possible. Since markets are complete, these trades are effectively possible. But, if only one trader has correct beliefs, then only one trader puts positive probability on the infinite sample paths that actually occur. So only this trader will have positive consumption, and thus positive wealth, in the limit.

For bounded complete market economies there is a survival index that determines which traders survive and which vanish. This index depends only on discount factors, the actual stochastic process of states, and, traders' beliefs about this stochastic process. Most importantly, for these economies, attitudes towards risk do not matter for survival. The literature also provides various results demonstrating how the market selects among learning rules. The market selects for traders who learn the true process over those who do not learn the truth, for Bayesians with the truth in the support of their prior over comparable non-Bayesians, and among Bayesians according to the dimension of the support of their prior (assuming that the truth is in the support).

In economies with incomplete markets, the market selection hypothesis can fail to be true. Blume and Easley (2006) show that if markets are incomplete, then rational traders may choose either savings rates or portfolio rules that are dominated by those selected by traders with incorrect beliefs. If some traders are irrationally optimistic about the payoff to assets, then the price of those assets may be high enough for rational traders to choose to consume more now, and less in the future. Their low savings rates are optimal, but as a result of their low savings rates the rational traders do not survive.

An alternative version of the market selection hypothesis is that asset markets select for traders with superior information. The research discussed above asks about selection over traders with different, but exogenously given, beliefs. Alternatively, if traders begin with a common prior and receive differential information they will have differing beliefs, but now they will care about each others' beliefs. In this case, the selection question is difficult because the information that traders have will be reflected in prices. If the economy is in a fully revealing rational expectations equilibrium, then there is no advantage to having superior information; see Grossman and Stiglitz (1980). So the question only makes sense in the more natural, but far more complex, case in which information is not fully revealed by market statistics. Figlewski (1978) shows that traders with information which is not correctly reflected in prices have an advantage in terms of expected wealth gain over those whose information is fully impounded in prices. But as expected wealth gain does not determine fitness this result does not fully answer the question. Mailath and Sandroni (2003) consider a Lucas trees economy with log utility traders and noise traders. They show that the quality of information affects survival, but so does the level of noise in the economy. Scuibba (2005) considers a Grossman and Stiglitz (1980) economy in which informed traders pay for information and shows that in this case uninformed traders do not vanish.

Conclusion

The modern literature has shown that the market selection hypothesis needs to be qualified. For some economies it acts much as the earlier writers conjectured; in others it does not select for profit maximizers or rational traders. Much work remains to be done, however. Blume and Easley (2006) and Sandroni (2000) mostly discuss selection in complete markets. Sandroni, though, points out that even when markets are incomplete, traders with log utility and rational expectations are favoured, while Blume and Easley construct some examples to show that the outcome of market selection can depend on market completeness. The connection between market structure and market selection is not well understood. The implications of market selection for asset pricing are known only for complete markets in the long run and some examples. Most economists' intuition about market behaviour and asset pricing comes from the study of market models that allow little or no agent heterogeneity. Taking heterogeneity seriously and chasing down its implications for market performance promises to be a rich area for future research.

M

See Also

- ▶ General Equilibrium
- ▶ Rational Expectations
- ▶ Rationality

References

Alchian, A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58: 211–221.

Amir, R., I. Evstigneev, T. Hens, and K.R. Schenk-Hoppe. 2005. Market selection and survival of investment strategies. *Journal of Mathematical Economics* 41: 105–122.

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

Blume, L., and D. Easley. 2002. Optimality and natural selection in markets. *Journal of Economic Theory* 107: 95–130.

Blume, L., and D. Easley. 2005. Rationality and selection in asset markets. In *The economy as an evolving*

complex system, ed. L. Blume and S. Durlauf. Oxford: Oxford University Press.
Blume, L., and D. Easley. 2006. If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets. *Econometrica* 74: 929–966.
Breiman, L. 1961. Optimal gambling systems for favorable games. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.
Cootner, P. 1967. *The random character of stock market prices*. Cambridge, MA: MIT Press.
DeLong, J.B., A. Shleifer, L. Summers, and R. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98: 703–738.
DeLong, J.B., A. Shleifer, L. Summers, and R. Waldmann. 1991. The survival of noise traders in financial markets. *Journal of Business* 64: 1–19.
Dutta, P., and R. Radner. 1999. Profit maximization and the market selection hypothesis. *Review of Economic Studies* 66: 769–798.
Enke, S. 1951. On maximizing profits: A distinction between Chamberlin and Robinson. *American Economic Review* 41: 566–578.
Evstigneev, I., T. Hens, and K.R. Schenk-Hoppe. 2006. Evolutionary stable stock markets. *Economic Theory* 27: 449–468.
Fama, E. 1965. The behavior of stock market prices. *Journal of Business* 38: 34–105.
Figlewski, S. 1978. Market 'efficiency' in a market with heterogeneous information. *Journal of Political Economy* 86: 581–597.
Friedman, M. 1953. *Essays in positive economics*. Chicago: University of Chicago Press.
Grossman, S.J., and J.E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.
Kelly, J.L. 1956. A new interpretation of information rate. *Bell System Technical Journal* 35: 917–926.
Koopmans, T. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.
Mailath, G., and A. Sandroni. 2003. Market selection and asymmetric information. *Review of Economic Studies* 70: 343–368.
Nelson, R., and S. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.
Sandroni, A. 2000. Do markets favor agents able to make accurate predictions? *Econometrica* 68: 1303–1342.
Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46: 55–67.
Scuibba, E. 2005. Asymmetric information and survival in financial markets. *Economic Theory* 25: 353–379.
Winter, S. 1964. Economic natural selection and the theory of the firm. *Yale Economic Essays* 4: 225–272.
Winter, S. 1971. Satisficing, selection and the innovating remnant. *Quarterly Journal of Economics* 85: 237–261.

Market Failure

John O. Ledyard

Abstract

Market failure occurs when there are too few markets, non-competitive behaviour, or non-existence, leading to inefficient allocations. Many suggested solutions for market failure, such as tax-subsidy schemes, property rights assignments, and special pricing arrangements, are simply devices for the creation of more markets. This remedy can be beneficial but, if the addition of markets creates either non-convexities or thin participation, then adding markets will simply lead to market failure from monopolistic behaviour. Examples are natural monopolies and informational monopolies. To achieve a more efficient allocation of resources in the presence of such fundamental failures one must explore non-market alternatives.

Keywords

Asymmetric information; Contingent claims markets; Free rider problem; Fundamental theorem of welfare economics; Increasing returns to scale; Lindahl prices; Market failure; Mechanism design; Monopoly; Monopsony; Natural monopoly; Non-competitive behaviour; Non-convexity; Non-existence of equilibrium; Pareto efficiency; Property rights reassignments; Rational expectations

JEL Classifications

D0

The best way to understand market failure is first to understand market success, the ability of a collection of idealized competitive markets to achieve an equilibrium allocation of resources that is Pareto optimal. This characteristic of markets, which was loosely conjectured by Adam Smith, has received its clearest expression in the theorems of modern welfare economics. For our

purposes, the first of these, named the first fundamental theorem of welfare economics, is of most interest. Simply stated it reads: (1) if there are enough markets, (2) if all consumers and producers behave competitively, and (3) if an equilibrium exists, then the allocation of resources in that equilibrium will be Pareto optimal (see Arrow 1951; Debreu 1959). Market failure is said to occur when the conclusion of this theorem is false; that is, when the allocations achieved with markets are not efficient.

Market failure is often the justification for political intervention in the marketplace (for one view, see Bator 1958, section V). The standard argument is that if market allocations are inefficient, everyone can and should be made better off. To understand the feasibility and desirability of such Pareto-improving interventions, we must achieve a deeper understanding of the sources of market failure. Since each must be due to the failure of at least one of the three conditions of the first theorem, we will consider those conditions one at a time.

The first condition requires there to be enough markets. Although there are no definitive guidelines as to what constitutes 'enough', the general principle is that if any actor in the economy cares about something that also involves an interaction with at least one other actor, then there should be a market for that something; it should have a price (Arrow 1969). This is true whether the something is consumption of bread, consumption of the smoke from a factory, or the amount of national defence. The first of these examples is a standard private good, the second is an externality, and the third is a public good. All need to be priced if we are to achieve a Pareto-optimal allocation of resources; without these markets, actors may be unable to inform others about mutually beneficial trades which can leave both better off.

The informational role of markets is clearly highlighted by a classic example of market failure analysed by Scitovsky (1954). In this example, a steel industry, which must decide now whether to operate, will be profitable if and only if a railway industry begins operations within five years. The railway industry will be profitable if and only if the steel industry is operating when the railway industry begins its own operations. Clearly each cares about the other and it is efficient for each to operate; the steel industry begins today and the railway industry begins later. Nevertheless, if there are only spot markets for steel, the railway industry cannot easily inform the steel industry of its interests through the marketplace. This inability to communicate desirable interactions and to coordinate timing is an example of market failure and has been used as a justification for public involvement in development efforts; a justification for national planning. However, if we correctly recognize that there are simply too few markets, we can easily find another solution by creating a futures market for steel. If the railway industry is able to pay today for delivery of steel at some specified date in the future then both steel and railway industries are able to make the other aware of their interests through the marketplace. It is easy to show that as long as agents behave competitively and equilibrium exists, the addition of futures markets will solve this type of market failure.

A completely different example of the informational role of markets arises when actors in the marketplace are asymmetrically informed about the true state of an uncertain world. The classic example involves securities markets where insiders may know something that outsiders do not. Even if it is important and potentially profitable for the uninformed actor to know the information held by the informed actor, there may not be enough markets to generate an efficient allocation of resources. To see this most clearly, suppose there are only two possible states of the world. Further, suppose there are two consumers, one of whom knows the true state and one of whom thinks each state is equally likely. If the only markets that exist are markets for physical commodities, then the equilibrium allocation will not in general be Pareto optimal. One solution is to create a contingent claims market. An 'insurance' contract can be created in which delivery and acceptance of a specified amount of the commodity is contingent on the true state of the world. Assuming both parties can, *ex post*, mutually verify which is indeed the true state of the world, if both behave competitively and an equilibrium

allocation exists, it will be Pareto optimal, given the information structure. A more general and precise version of this theorem can be found in Radner (1968).

Analysing this example further we note that in equilibrium the prices of commodities in the state that is not true will be close to or equal to zero, since at positive prices the informed actor will always be willing to supply an infinite amount contingent on the false state, knowing delivery will be unnecessary. If the uninformed actor is clever and realizes that prices will behave this way in equilibrium then he can become informed simply by observing which contingency prices are zero. If he then uses this information, which has been freely provided by the market, the equilibrium will be Pareto optimal under full information. In a very simple form, this is the idea behind rational expectations (see Muth 1961). With clever competitive actors, it may not be necessary to create all markets in order to achieve a Pareto-efficient equilibrium allocation.

Completing markets seems to be an easy technique to correct market failure. The suggestions that taxes and subsidies (Pigou 1932) or property rights reassignments (Coase 1960) can cure market failure follow directly from this observation. However, an unintended consequence can sometimes occur after the creation of these markets. In some cases, adding more markets may cause conditions (2) and (3) of the first theorem to be false. Curing one form of market failure can lead to another. To understand how this happens and how the second condition requiring competitive behaviour can be affected, consider the informed consumer in our previous example. If he realizes that the uninformed consumer is going to make inferences based indirectly on his actions then he should not behave competitively because he could do better by pretending to be uninformed. He can, by strategically limiting the supply of information of which he is the monopoly holder, do better than if he behaved competitively. It is only his willingness to supply infinite amounts of the commodity in the false state that gives away his knowledge. Supplying only a little commodity contingent on that (false) state in return for a small payment today would not allow the uninformed agent to infer anything and would allow the informed agent to make a profit from his monopoly position. This is not very different from the standard example of a violation of condition (2), monopoly supply of a commodity.

A different example of this phenomenon of unintended outcomes arises when markets are created to allocate public goods. It is now well known that the introduction of personal, Lindahl prices to price individual demands for a public good does indeed lead to Pareto-optimal allocations if consumers behave competitively (see Foley 1970). However, under this scheme, each agent becomes a monopsonist in one of the created markets and, therefore, has an incentive to understate demand and not to take prices as given. This is the phenomenon of 'free riding', often alluded to as the reason why the creation of markets may not be a viable solution to market failure. To understand why, let us now examine the second condition of the first theorem in more detail.

The second condition of the first theorem about market success is that all actors in the marketplace behave competitively. This means that each must act as if they cannot affect prices and, given prices, as if they follow optimizing behaviour. Consumers maximize preferences subject to budget constraints and producers maximize profits, each taking prices as fixed parameters. This condition will be violated when actors can affect the values that equilibrium prices take and in so doing be better off. The standard example of market failure due to a violation of this condition is monopoly, in which one actor is the sole supplier of an output. By artificially restricting supply, this actor can cause higher prices and make himself or herself better off even though the resulting equilibrium allocation will be inefficient.

Can we correct market failure due to non-competitive behaviour? To find an answer let us first isolate those conditions under which agents find it in their interests to follow competitive behaviour. The work of Roberts and Postlewaite (1976) has established that if each agent holds only a small amount of resources relative to the aggregate available, then they will usually be unable to manipulate prices in any significant way and will act as price takers. It is the depth of

the market that is important. This is also true when the commodity is information. If each agent is informationally small, in the sense that he either knows very little or what he does know is of little importance to others, then he loses little by behaving competitively (see Postlewaite and Schmeidler 1986). On the other hand, if he is informationally important, as in the earlier example, he may have an incentive to behave non-competitively. The key is the size of the agent's resources, both real and informational, relative to the market.

The solution to market failure from non-competitive behaviour then seems to be to ensure that all agents are both resource and informationally small. Of course this must be accomplished through direct intervention as in the antitrust laws and the securities market regulations of the United States and may not be feasible. For example, it may not be possible to correct this type of market failure by simply telling agents to behave competitively. In such an attempt, one would try to enforce a public policy that all firms must charge prices equal to the marginal cost of output. But, unless the costs and production technology of the firm can be directly monitored, a monopolist can easily act as if he were setting price equal to marginal cost while using a false cost curve. It would be impossible for an outside observer to distinguish this non-competitive behaviour from competitive behaviour without directly monitoring the cost curve. If the monopolist were a consumer whose preferences were unobservable, then even monitoring would not help. In general, market failure from non-competitive behaviour is difficult to correct while still retaining markets. We will hint at some alternatives below.

Expansion of the number of markets can also lead to violations of the third condition of the first theorem. For illustration we consider three examples. The first and simplest of these is the case of increasing returns to scale in production. The classic case is a product that requires a fixed set-up cost and a constant marginal cost to produce. (More generally we could consider non-convex production possibilities sets.) If the firm acts competitively in this industry and if the price is above marginal cost the firm will supply an infinite amount. If the price is at or below marginal cost the firm will produce nothing. If the consumers' quantity demand is positive and finite at a price equal to marginal cost, then there is no price such that supply equals demand. Equilibrium does not exist. The real implication of this situation is not that markets do not equilibrate or that trade does not take place, it is that a natural monopoly exists. There is room for at most one efficient firm in this industry. Again it is the assumption of competitive behaviour that is ultimately violated.

The next example, due to Starrett (1972), involves an external diseconomy. Suppose there is an upstream firm that pollutes the water and a downstream firm that requires clean water as an input into its production process. It is easy to show that if such a diseconomy exists and if the downstream firm always has the option of inaction (that is, it can use no inputs to produce no outputs at zero cost), then the aggregate production possibilities set of the economy when expanded to allow enough markets cannot be convex (see Ledyard 1976 for a formal proof). If the production possibilities set of the economy is non-convex, then, as in the last example, it is possible that a competitive equilibrium will not exist. Expansion of the number of markets to solve the inefficiencies due to external diseconomies can lead to a situation in which there is no competitive equilibrium.

The last example, first observed by Green (1977) and Kreps (1977), arises in situations of asymmetric information. Recall the earlier example in which one agent was fully informed about the state of the world while the other thought each state was equally likely. Suppose preferences and endowments in each state are such that if both know the state then the equilibrium prices in each state are the same. Further, suppose that if the uninformed agent makes no inferences about the state from the other's behaviour then there will be different prices in each state. Then no (rational expectations) equilibrium will exist. If the informed agent tries to make inferences the prices will not inform him, and if the uninformed agent does not try to make inferences the prices will inform him. Further, it is fairly easy to show that if a market for information could be created (ignoring incentives to hide information) the

resulting possibilities set is in general non-convex. In either case there is no equilibrium.

Most examples of non-existence of equilibrium seem to lead inevitably to non- competitive behaviour. In our example of non-existence due to informational asymmetries, it is natural for the informed agent to behave as a monopolist with respect to that information. In the example of the diseconomy, if a market is created between the upstream and the downstream firm, each becomes a monopoly. If there is a single polluter and many pollutees, the polluter holds a position similar to a monopsony. The non-existence problem due to the fundamental non-convexity caused by the use of markets to eliminate external diseconomies is simply finessed by one or more of the participants assuming non-competitive behaviour. An outcome occurs but it is not competitive and, therefore, not efficient.

Market failure, the inefficient allocation of resources with markets, can occur if there are too few markets, non-competitive behaviour, or non-existence problems. Many suggested solutions for market failure, such as tax-subsidy schemes, property rights assignments, and special pricing arrangements, are simply devices for the creation of more markets. If this can be done in a way that avoids non-convexities and ensures depth of participation, then the remedy can be beneficial and the new allocation should be efficient. On the other hand, if the addition of markets creates either non-convexities or shallow participation, then attempts to cure market failure from too few markets will simply lead to market failure from monopolistic behaviour. Market failure in this latter situation is fundamental. Examples are natural monopolies, external diseconomies, public goods and informational monopolies. If one wants to achieve efficient allocations of resources in the presence of such fundamental failures one must accept self-interested behaviour and explore non-market alternatives. A literature using this approach, sometimes called implementation theory and sometimes called mechanism design theory, was initiated by Hurwicz (1972) and is surveyed in Groves and Ledyard (1986). More recent results can be found at mechanism design and mechanism design (new developments).

See Also

- ▶ Incentive Compatibility
- ▶ Incomplete Contracts
- ▶ Incomplete Markets
- ▶ Mechanism Design
- ▶ Mechanism Design (New Developments)
- ▶ Pareto Efficiency
- ▶ Welfare Economics

Bibliography

Arrow, K. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K. 1969. The organization of economic activity: Issues pertinent to the choice of market versus non-market allocation. In Joint Economic Committee, *The Analysis and Evaluation of Public Expenditures: The PPB System*. Washington, DC: Government Printing Office.

Bator, F. 1958. The anatomy of market failure. *Quarterly Journal of Economics* 72: 351–379.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. Cowles Foundation Monograph No. 17. New York: Wiley.

Foley, D. 1970. Lindahl's solution and the core of an economy with public goods. *Econometrica* 38: 66–72.

Green, J. 1977. The nonexistence of informational equilibria. *Review of Economic Studies* 44: 451–463.

Groves, T., and J. Ledyard. 1986. Incentive compatibility ten years later. In *Information, incentives, and economic mechanisms*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization*, ed. C.B. McGuire and R. Radner. Amsterdam: North-Holland.

Kreps, D. 1977. A note on 'fullfilled expectations' equilibria. *Journal of Economic Theory* 14: 32–43.

Ledyard, J. 1976. Discussion of 'on the nature of externalities'. In *Theory and measurement of economic externalities*, ed. S. Lin. New York: Academic Press.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Pigou, A. 1932. *The economics of welfare*. 4th ed. New York: Macmillan.

Postlewaite, A., and D. Schmeidler. 1986. Differential information and strategic behavior in economic environments: A general equilibrium approach. In *Information, incentives, and economic mechanisms*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36: 31–58.
Roberts, J., and A. Postlewaite. 1976. The incentives for price-taking behavior in large exchange economies. *Econometrica* 44: 115–127.
Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62: 70–82.
Starrett, D. 1972. Fundamental non-convexities in the theory of externalities. *Journal of Economic Theory* 4: 180–199.

Market Institutions

John McMillan

Abstract

Market-supporting institutions ensure that property rights are respected, that people can be trusted to live up to their promises, that externalities are held in check, that competition is fostered, and that information flows smoothly. Evidence is reviewed here on some market institutions: property rights and contracting with and without the law, and mechanisms to sustain information flow in markets.

Keywords

Adverse selection; Akerlof, G.; Asymmetric information; Coase, R.; Contracting; Equity market; Fisheries; Grameen Bank; Group lending; India; Information; Information transmission; Land titles; Market institutions; Markets for lemons; Micro-credit; Moral hazard; Non-price information; North, D.; Property rights; Screening; Signalling; Trade credit; Transaction costs

JEL Classifications

O1

In order to work as they should, markets need institutions. Defining the rules of the game, institutions consist of the constraints, formal and informal, on economic and political actors (North 1991). Market institutions serve to limit transaction costs: the time and money spent locating trading partners, comparing their prices, evaluating the quality of the goods for sale, negotiating agreements, monitoring performance and settling disputes (McMillan 2002).

The notion that institutions matter is as old as the study of economics. For markets to create gains from trade, as Adam Smith recognized, the state must define property rights and enforce contracts.

That institutions matter is also one of the chief insights from modern economics. In the presence of informational asymmetries, markets can falter. If buyer and seller have different information about the item to be exchanged, a 'lemons market' may arise. Unable to distinguish high-quality goods, buyers may be unwilling to pay a price that elicits supply of anything other than low-quality items. Potential gains from trade go unrealized (Akerlof 1970). When information is distributed unevenly – as is ubiquitous in the real world of economics, even if most of the textbooks have yet to bring it on board – prices do not incorporate all relevant information, and so non-price information is needed (Spence 1973; Rothschild and Stiglitz 1976). Limiting the inefficiencies from informational asymmetries requires mechanisms for signalling and screening: devices like reputation, warranties and credentials, as well as in some cases government-set rules and regulations. A more nuanced view of market processes is called for than the institution-free textbook account of price equilibration via supply and demand.

Evidence on the role of market-supporting institutions is accumulating. Much of the evidence comes from developing countries and countries in transition from communist central planning. Where markets work smoothly, in affluent countries, the market-supporting institutions are almost invisible. It is hard to find evidence of lemons markets in a country like the United States, because institutional solutions have evolved. By contrast, where markets work badly, in poor countries, the absence of institutions is conspicuous (Klitgaard 1991). A few examples are given in what follows.

Property Rights and Contracting

Institutional innovation sometimes occurs even in affluent countries. An experiment in property rights has arisen in fisheries. Worldwide, fisheries are in crisis. Overfishing results from an externality: the costs of any one fisher's taking too many fish are mostly borne by others. Applying the idea of Ronald Coase (1960) of defining property rights to solve an externality, the New Zealand government has created, essentially, property rights in the fish. Fishers are assigned quotas that define, by species, their allowable fish catch. The quotas are tradable, so they end up with those fishers with the highest willingness to pay, which probably leads to an efficient allocation. Property rights in fish do not come for free, however, but require extensive, costly government monitoring (Grafton et al. 2000). Military aircraft patrol the oceans. Each step of every single fish's journey from landing to final sale is documented, with catch reports, buyers' receipts, cold-storage records and export invoices being collated. Fishery inspectors police breaches. The costs of overseeing the quotas have yielded a return, as fish stocks have been successfully conserved.

Another property-rights experiment has occurred in residential land. In cities in every developing country there are squatters, poor people living on land to which they hold no legal rights. Ad hoc property rights exist even in the absence of formal legal protections, as neighbourhood associations and the squatters themselves guard the land. However, the inability to appeal to the law brings some inefficiencies. Hernando de Soto (2000) argued that, if the impoverished squatters held land titles, they would acquire access to capital markets, because they would then have collateral to offer. In Peru, following de Soto's advocacy, over a million squatter households were granted title to the land they occupied. The effects of this huge inauguration of property rights showed up, unexpectedly, not in the capital market but in the labour market. Householders' borrowing increased little, but hours worked outside the home by adult household members increased and hours worked by their children decreased (Field 2003; Field and Torero 2004). Without land titles, householders stayed at home to watch over their property, sending their children out to work. Holding land titles, they felt secure enough to enter the workforce. Establishing the market institution brought instant welfare gains. However, the gains came in an unforeseen form, illustrating the difficulty in general of anticipating the effects of institutional reform (McMillan 2004).

With contracting, as with property rights, informal substitutes operate in the absence of formal institutions. Small firms make deals with each other and get finance, using personal networks and ongoing relationships to substitute for missing laws of contract and using retained earnings and trade credit to make up for a lack of access to financial markets (Fafchamps 2004; McMillan and Woodruff 2002). Large firms also can prosper without institutions, coping instead by cultivating favours from politicians. Where the lack of institutions shows up is for small firms wishing to grow. Needing to make large, discrete investments, they can no longer rely on retained earnings and trade credit, so they may be unable to grow if the financial market is underdeveloped. Needing to deal with increasing numbers of trading partners, they cannot continue to rely on personal connections but must start to use the law of contract. The firm-size distribution in a typical developing country shows a missing middle, with a lot of employment in tiny firms and quite a lot in large firms, but not much in mid-sized firms (Snodgrass and Biggs 1996). The missing middle is a symptom of weak legal and regulatory institutions.

Information Transmission

An archetypical lemons market existed in India in the 1970s (Klitgaard 1991). Quality fresh milk was hard to find because vendors routinely watered it down. Buyers could not assess the milk's butterfat content, and so the low-quality milk drove out the high-quality milk. Launching a campaign against adulterated milk, the National Dairy Development Board provided inexpensive machines to measure butterfat content as the milk

moved from farmer to wholesaler to vendor. It also set up payment schemes making the price of milk reflect its measured quality and created brand names to give buyers trust in what they were getting. As a result of this coordinated initiative, quality improved and consumption rose.

The loan market is impeded by information asymmetries: both adverse selection (a lender may find it hard to distinguish whether any given loan applicant is a good credit risk) and moral hazard (a borrower, having received a loan, may have an incentive to default). Since these transaction costs are proportionately larger for small than for large loans, small lenders often pay exorbitant interest rates or are frozen out of the loan market. In Bangladesh's Grameen Bank and other microcredit banks, tiny loans are made to poor people via groups of borrowers. Each group member is held responsible for any other member's loan. Being neighbours, the group members know each other's business better than any banker, can monitor each other's use of the loans and can invoke social sanctions to discipline defaulters. Group lending is an elegant solution to the loan market's informational asymmetries.

The equity market relies heavily on institutions. For shareholders, who lack information about the firm's affairs, evaluating managers is difficult, and so a lemons market may arise. In many countries, lax oversights allow controlling shareholders to expropriate minority shareholders (Johnson et al. 2000). If the rules governing the financial markets are inadequate, investors are reluctant to buy stocks because they are unwilling to trust managers, and so firms do not get the finance they need. A well-functioning equity market relies on a complex set of interrelated institutions, formal and informal, to foster information flow (Black 2001). First, reputations for honest dealings must be built up by auditors, law firms, investment banks and the business press. Second, there are self-regulating private-sector bodies such as industry associations as well as the stock exchange, with its rules on listing firms' financial reporting and its sanction of delisting. Third, the equity market rests on state-provided mechanisms: not only laws requiring that investors receive accurate data, but also an activist regulator. The law's transaction costs (Glaeser and Shleifer 2003) mean that a regulator supplements the courts in setting and enforcing the rules of the game.

Conclusion

Market-supporting institutions ensure that property rights are respected, that people can be trusted to live up to their promises, that externalities are held in check, that competition is fostered and that information flows smoothly (McMillan 2002). Without institutions, the promise of efficient markets goes unrealized.

See Also

- ▶ Akerlof, George Arthur (Born 1940)
- ▶ Arrow, Kenneth Joseph (Born 1921)
- ▶ Development Economics
- ▶ Growth and Institutions
- ▶ Institutionalism, Old
- ▶ Microcredit
- ▶ Property Law, Economics And
- ▶ Search Theory (New Perspectives)
- ▶ Spence, A. Michael (Born 1943)
- ▶ Stiglitz, Joseph E. (Born 1943)

Bibliography

Akerlof, G. 1970. The market for 'lemons'. *Quarterly Journal of Economics* 84: 488–500.

Black, B. 2001. The legal and institutional preconditions for strong stock markets. *UCLA Law Review* 48: 781–855.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

de Soto, Hernando. 2000. *The mystery of capital*. New York: Basic Books.

Fafchamps, M. 2004. *Market institutions in Sub-Saharan Africa*. Cambridge, MA: MIT Press.

Field, E. 2003. *Entitled to work*. Unpublished manuscript. Cambridge, MA: Harvard University.

Field, E., and M. Torero. 2004. *Do property titles increase credit access among the urban poor?* Unpublished manuscript. Cambridge, MA: Harvard University.

Glaeser, E., and A. Shleifer. 2003. The rise of the regulatory state. *Journal of Economic Literature* 41: 401–425.

Grafton, R., D. Squires, and K.J. Fox. 2000. Private property and economic efficiency. *Journal of Law and Economics* 43: 679–714.
Johnson, S., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2000. Tunneling. *American Economic Review Papers and Proceedings* 90: 22–7.
Klitgaard, R. 1991. *Adjusting to reality*. San Francisco: ICS Press.
McMillan, J. 2002. *Reinventing the Bazaar*. New York: Norton.
McMillan, J. 2004. Avoid hubris. *Finance and Development* 41: 34–37.
McMillan, J., and C. Woodruff. 2002. The central role of entrepreneurs in transition economies. *Journal of Economic Perspectives* 16: 153–170.
North, D. 1991. Institutions. *Journal of Economic Perspectives* 5: 97–112.
Rothschild, M., and J.E. Stiglitz. 1976. Equilibrium in competitive insurance markets. *Quarterly Journal of Economics* 90: 629–650.
Snodgrass, D., and T. Biggs. 1996. *Industrialization and the small firm*. San Francisco: ICS Press.
Spence, A. 1973. Job market signaling. *Quarterly Journal of Economics* 87: 355–374.

Market Microstructure

Maureen O'Hara

Abstract

Market microstructure research uses the rules and trading protocols of markets to analyse price formation in asset markets. Microstructure research shows how markets provide liquidity and price discovery, and how prices come to reflect information. Of particular importance to this process is how market participants learn from market data. Microstructure researchers often consider issues related to market structure, in particular how changing features of the market affect the price process. Empirical microstructure research uses high-frequency data-sets, and develops statistical approaches to deal with such data.

Keywords

Adverse selection; Asset pricing; Asymmetric information; Auto-conditional duration model; Bayes' rule; Bid–ask markets; Bonds; Censored sampling; Electronic commerce; Market microstructure; Inventory models; Inventory risk; Liquidity; Market structure; Noise traders; Options; Price discovery; Rational expectations; Sequential models; Spreads

JEL Classifications

G10; G34

Market microstructure studies the behaviour and formation of prices in asset markets. Whereas economic analyses of price formation generally abstract from any particular price-setting mechanisms, market microstructure relies on the specific rules and protocols of markets to analyse how prices are determined. This focus on the microstructure of the market provides insights into how the design of markets affects the price process, detailing both how individual prices are determined and how those prices evolve over time. Such insights are useful for a wide range of issues in asset pricing, as well as for guiding econometric investigations of high frequency data. In addition, microstructure research analyses structural issues in securities trading, such as the role and function of exchanges, the optimal design of trading systems, and the optimal regulation of securities markets.

Fundamental to microstructure research is the realization that asset prices are set in actual markets, and not by fictional auctioneers. Thus, while the forces of supply and demand ultimately underlie all asset prices, the specific formation and evolution of prices is much more complex. Buyers and sellers, for example, need not arrive synchronously, making the determination of a market-clearing price at a point in time problematic. When traders do arrive at markets, they may also face a range of market frictions such as transactions costs, search costs and the like (see Stoll 2001). Furthermore, the value of assets may change over time, with some traders potentially knowing more about future values than other traders. Markets facilitate the trading of assets by providing liquidity and price discovery, and how they do so depends on the rules and structure of the market (see O'Hara 2003).

Canonical Models in Microstructure

Early microstructure models focused on the specific market structure found in organized stock markets. In such markets, a designated market-maker or specialist quotes prices to buy or sell units of the asset. By serving as counter-party to buyers and sellers, the market-maker solves the asynchronicity problem noted above by standing ready to provide liquidity on either side of the market. The market-maker earns the 'spread', or the difference between the price at which he buys shares (the bid) and the price at which he will sell shares (the ask). In return, however, the market-maker has to bear inventory risk, essentially going long when traders wish to sell, and short when traders wish to buy.

There is an extensive literature analysing the market-maker's pricing problem in the presence of inventory risk (for a review of models, see O'Hara 1995). In general, such models assume risk-averse market-makers facing exogenous holding costs in a setting in which all agents are symmetrically informed and 'true' asset prices are assumed fixed or, at least, stationary processes. An important feature of the equilibrium is that there is no single price: the price the market-maker sets depends upon whether the trader wishes to buy or sell, and on how much he wishes to trade. Prices change over time in response to the specialist's inventory position, his market power and parameters relating to the supply and demand for the asset. Such inventory models have been extended to a wide variety of market settings such as foreign exchange, bond markets, and options and futures markets. Empirical analyses find substantial support for the predictions of inventory models.

An alternative class of microstructure models considers price-setting when some agents have better information about the asset's true value than do other agents. The impetus for such models was an early paper by Treynor (1971), who noted that traders arriving at the market included those who needed to trade for liquidity reasons, those with better information about the asset's true value, and those who thought they had better information but were in fact incorrect. Treynor conjectured that the market-maker's prices were a balancing act offsetting his losses to the informed traders with his gains from the liquidity and noise traders. Viewed from this perspective, a spread arises naturally in security markets, independent of any inventory or transactions costs explanations. Fisher Black (1986) expanded on this notion to highlight the important role played by noise or liquidity traders in allowing markets to become efficient.

An intriguing implication of this research is that, if some traders do have better information about the asset's true value, then the nature of the order flow can be informative as to future asset values. Consequently, the market-maker's price-setting problem evolves from being a simple balancing of expected gains and losses to that of learning how to extract information from the order flow. With the market-maker drawing inferences from the order flow, this sets the stage for traders to consider the impact of their trades as well, particularly if they are attempting to profit on private information.

There are two general approaches to modelling price-setting in the presence of asymmetric information, sequential trade models and Kyle (1985) models. Glosten and Milgrom (1985) consider a risk-neutral market-maker facing known populations of informed and uninformed traders, where traders arrive sequentially to the market. The market-maker knows these population parameters, but does not know the identity of any individual trader. The market-maker does know, however, that traders informed of good news will all want to buy, while those informed of bad news will all want to sell. Consequently, the market-maker's conditional expectation of the asset's value also differs with trade direction, and it is these conditional expectations that become his bid and ask prices. Based on the trade that actually occurs, the market-maker updates his beliefs regarding the asset's value using Bayes' rule. The continued one-sided trading of the informed traders eventually forces prices to the true equilibrium level.

Sequential trade models provide an elegant means to characterize the relation between trades and prices on a tick-by-tick basis. Because the market-maker learns from trades, the evolution of prices depends on the order flow, as does the size and movement of the spread. More complex

analyses demonstrate a role for other market information in affecting price behaviour. Trade size, for example, may be informative as informed traders prefer to trade larger rather than smaller amounts (Easley and O'Hara, 1987). The time between trades may also have information content as a signal of the existence of new information, and this, in turn, can impart information content to volume (Easley and O'Hara 1992). Trade location, trade in correlated assets, and alternative order types can also have information content. Because of their tick-by-tick focus, sequential models are particularly useful for guiding empirical analysis of microstructure data, an issue we will return to shortly.

An alternative modelling approach is a Kyle (1985) model. Kyle models focus on the dual problems facing the market-maker, who must figure out what the informed traders know, and the informed trader, who wishes to exploit his private information for profit. The Kyle model uses a batch-auction framework in which the market-maker sees the aggregated trades of both the informed traders and the noise or liquidity traders, and based on this order flow he sets a single price. The market-maker conjectures a trading strategy for the informed trader that is linear in the asset's true value, while the informed trader conjectures a pricing strategy for the market-maker that is linear in the total order flow. In equilibrium, both conjectures must be correct, a feature typical of rational expectations equilibrium models. As in sequential trade models, the market-maker's price reflects his conditional expected value for the asset, this conditional expectation changes as he learns from the order flow, and prices eventually adjust to true values. Back and Baruch (2004) demonstrate conditions under which the Kyle and Glosten–Milgrom models essentially converge.

An important feature of Kyle models is their ability to characterize the trading strategy of the informed trader. The optimal strategy for the informed trader is essentially to hide his trades in the noise trade, and he varies his trades over time in response to the market-maker's growing precision of his beliefs about the asset's true value. Holden and Subrahmanyam (1992) show that, if there are many informed traders, then their combined trading actions force prices almost instantaneously to true values, a result again reminiscent of rational expectations models. A wide range of research has considered variants of the Kyle model allowing for different types of information structures, for uninformed traders to also act strategically, and for the market-maker to have differential information.

These two asymmetric information-based modelling approaches allow researchers to address a broad range of issues in the trading of financial assets, and are particularly useful in demonstrating how markets perform their price discovery function. Because market-makers are risk neutral and unconstrained as to their inventory holdings, liquidity issues in these models reflect more difficulties induced by the potential information content of trades, rather than the risk-bearing considerations that arise in inventory models. As both effects are likely to be present in actual markets, a wide range of research has investigated empirically how spreads and price changes are influenced by information, inventory, and the fixed costs of making markets.

Research Directions in Microstructure Research

The growth of financial asset markets worldwide, as well as the increasing availability of high frequency microstructure data from a wide array of markets, has allowed microstructure researchers to investigate a broad range of issues, both empirical and theoretical. I highlight here a few areas that are of particular importance.

Econometrics of High-frequency Data

Microstructure data allows researchers to analyse the evolution of prices and market data on a second-by-second basis. Indeed, most microstructure data sets include millions of observations, raising a range of econometric issues. Of particular importance are the periodicity of the data, biases introduced by market structure protocols, optimal statistical models for evaluating the behaviour of prices and spreads, and data sampling issues. Hasbrouck (2006) discusses each of these topics.

Because prices arise only when there are trades, price data is not spaced uniformly throughout the trading day. This introduces a censored sampling problem as prices can be thought of as draws from the true asset value distribution, but where the timing of the draws may not be independent of evolution of the value process itself. Engle and Russell (1998) exploit this insight to develop the auto-conditional duration (ACD) model to analyse the evolution of intra-day volatility. A related problem is sampling across assets, as non-synchronicity of trading may result in price observations that lag true value innovations across stocks. A number of authors have considered the implications of non-synchronous trading for cross-sectional econometric analyses.

A variety of authors also consider the time-series properties of microstructure data, with a particular focus on decomposing price movements into those associated with the value process and those reflecting noise arising from the microstructure such as tick size constraints, bid/ask bounce, price continuity rules, and so on. These econometric issues are particularly important for asset pricing research.

Asset Pricing – Liquidity and Information Risk

Microstructure models analyse the liquidity and price discovery roles markets play in asset pricing. Recent research has focused on whether these two market roles also affect asset returns. Amihud and Mendelson (1986) first suggested that liquidity could influence asset returns by affecting an investor's overall cost of trading.

Numerous empirical researchers have investigated whether spreads, a proxy for these liquidity costs, are related to asset returns, but the empirical evidence has been mixed. More recent research by Pastor and Stambaugh (2004) using lagged volume measures of liquidity provides stronger evidence, and the authors propose a liquidity factor to explain asset returns. One reason why this effect may arise is commonality in liquidity. Chordia et al. (2000) find that liquidity measures appear to vary systematically across stocks, and these effects may be time-varying. Other researchers have found similar commonality effects in bond market liquidity measures.

A second research stream considers the price discovery process, and whether investors require higher returns to hold stocks for which a greater fraction of the available information is private rather than public. Easley et al. (2002) derive measures of information-based trading using a structural microstructure model, and demonstrate that asset returns are explained by these information measures. What generates this effect is the inability to diversify optimally, as uninformed traders always lose to informed traders, who are better able to shift their portfolio weights to reflect true values. Empirical research supports a distinct role for both liquidity and information risk in affecting asset returns.

Electronic Markets and Trading Systems

Microstructure models have typically analysed price-setting on a centralized market with a designated market-maker (or makers). While such a setting corresponds well to an exchange or dealer market, it is less applicable to the wide variety of electronic markets now used to trade many financial assets. Of particular importance are electronic trading systems which rely on the aggregation of limit orders to effectuate trades. Orders to buy and sell at a specific price and quantity are collected in the 'book', with price and time priority rules dictating how such orders are handled. At any point in time, a spread exists between the highest (lowest) price at which someone is willing to buy (sell) the asset. In such systems, trades arise when orders cross, imparting an importance to the order decisions of individual traders.

Traders face complex decision problems in placing orders due to the uncertainty of execution of any order. Of particular concern is that uninformed traders may face an adverse selection problem in that their trades are more likely to execute when there is new information, causing them to buy when there is bad news and sell when there is good news. This difficulty is further compounded by trading protocols that allow limit orders to 'sweep the book' and thereby trigger the execution of many individual orders as the opposite side of a large order. There is a substantial literature looking at the behaviour of such electronic markets, but the complexity of these markets leaves many important issues yet to be resolved.

Market Structure

Microstructure research is traditionally concerned with issues related to the design and structure of markets. The rise of new markets and trading technologies has raised a plethora of market structure issues. Of particular interest to many researchers are questions relating to transparency, or what information is available to traders and when can they see it. Bond markets, for example, were traditionally opaque, but new reporting rules have increased their transparency. Numerous authors have investigated how this has changed the liquidity and efficiency of the bond market. Option markets traditionally faced little competition, but the development of a national options market in the United States, along with the rise of electronic competitors, has changed this market structure. Regulatory changes in the United States and Europe have also dramatically affected market structure in equities, raising questions as to the efficacy of these new rules. Finally, the markets themselves are evolving from member-owned cooperatives to publicly traded firms, raising a host of issues relating to corporate governance and self-regulation. Microstructure research provides a means to evaluate the economic impact of these changes and to suggest alternative structures for the trading of financial assets.

See Also

- ▶ Adverse Selection
- ▶ Noise Traders

Bibliography

Amihud, Y., and H. Mendelson. 1986. Asset pricing and the bid-ask spread. *Journal of Financial Economics* 17: 223–249.

Back, K., and S. Baruch. 2004. Information in securities markets: Kyle meets Glosten and Milgrom. *Econometrica* 72: 433–465.

Black, F. 1986. Noise. *Journal of Finance* 41: 529–543.

Chordia, T., R. Roll, and A. Subrahmanyam. 2000. Commonality in liquidity. *Journal of Financial Economics* 56: 3–28.

Easley, D., and M. O'Hara. 1987. Price, trade size, and information in securities markets. *Journal of Financial Economics* 19: 69–90.

Easley, D., and M. O'Hara. 1992. Time and the process of security price adjustment. *Journal of Finance* 47: 577–605.

Easley, D., S. Hvidkjaer, and M. O'Hara. 2002. Is information risk a determinant of asset prices? *Journal of Finance* 56: 2185–2221.

Engle, R., and J.R. Russell. 1998. Autoregressive conditional duration: a new model for irregularly spaced data. *Econometrica* 66: 1127–1162.

Glosten, L., and P. Milgrom. 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14: 71–100.

Hasbrouck, J. 2006. *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. New York: Oxford Economic Press.

Holden, C.W., and A. Subrahmanyam. 1992. Long-lived private information and imperfect competition. *Journal of Finance* 47: 247–270.

Kyle, A. 1985. Continuous auctions and insider trading. *Econometrica* 53: 1315–1335.

O'Hara, M. 1995. *Market Microstructure Theory*. Boston: Blackwell.

O'Hara, M. 2003. Presidential address: Liquidity and price discovery. *Journal of Finance* 58: 1335–1354.

Pastor, L., and R. Stambaugh. 2004. Liquidity risk and expected stock returns. *Journal of Political Economy* 111: 642–685.

Stoll, H. 2001. Market frictions. *Journal of Finance* 55: 1479–1514.

Treynor, J. 1971. The only game in town. *Financial Analysts Journal* 27(12–14): 22.

Market Period

D. R. Helm

Keywords

Expectations; Imperfect competition; IS–LM; Long run and short run; Market period; New classic macroeconomics; Perfect foresight; Perfect markets; Rational expectations; Temporary equilibria

JEL Classifications

D4

The concept of market period was introduced by Marshall to define markets according to the time period over which they extended. It was thus an additional classification of markets to that of location or space (*Principles*, V.i.6). This distinction became the modern textbook one between the short period and the long, reducing Marshall's more complex three-period classification. As he put it,

> we shall find that if the period is short, the supply is limited to the stores which happen to be at hand; if the period is long, the supply will be influenced, more or less, by the cost of producing the commodity in question; and if the period is very long, this cost will in its turn be influenced, more or less, by the cost of producing the labour and material things required for producing the commodity.

Hence the short run is that period for which stocks are constant, the long run that period where price is determined by the costs of production (but factors are constant) and the very long run that period where all factors vary.

The Marshallian market period was, as Hicks pointed out (1965, chapter 5) one of the ways in which Marshall used his 'static method'. For in the short period, Hicks goes on to say, Marshall could treat the industry as if it were in static equilibrium. Capital, fixed in the short period, is like land in Ricardo, and it earns a rent. In the longer run, the static method breaks down, as capital becomes variable, like labour.

The concept of Marshallian short period has been used extensively in the theory of the firm, in terms of short- and long-run equilibria, and the defining of cost curves according to this classification. Harrod, in 1934, linked this Marshallian concept with the new theory of imperfect competition developed by Joan Robinson (and Chamberlin), to look at the process of imperfect competition and the impact of entry on short- and long-run profit maximization.

The Marshallian short run concept was taken over into macroeconomics by Keynes as one of three components of Marshall's theory which he used to construct the *General Theory* (the others were partial equilibrium and thus exogenous expectations, and the representative firm aggregate which Keynes took over as the economy). However the Keynesian use of market period was not universally adopted in macroeconomics and it is Hicks's much more restrictive concept used in the IS–LM framework which is now much more familiar. The Hicksian 'week' is a market period in which fundamentally it is stocks that are constant, while the Keynesian 'year' allows for an element of 'user cost' whereby the utilization of capital affects the future demand for capital.

The Hicksian week and the concept of temporary equilibrium associated with it were first set out by Hicks in the middle chapters of *Value and Capital* in 1939, and were subsequently revised in an important, somewhat neglected essay entitled 'Methods of Dynamic Analysis' in 1956, reprinted in *Money, Interest and Wages* (1982). These concepts formed part of an attempt to construct a theory of dynamics, going beyond Marshall's static analysis. The alternative extreme hypothesis of allowing all factors to vary is a longer-run theory, and forms the basis of general equilibrium and growth theory. Both the Hicksian and Keynesian theories attempt to construct an intermediary period, and for each the corresponding problems were to decide which factors are to be allowed to vary, which to stay constant and what process of adjustment to vary, which to be employed by firms. But once these theoretical assumptions have been made, the individual periods become discrete rather than continuous (as in the longer-run case). Thus a theory of dynamics based on the Hicksian week requires an additional theory by which to link the discrete periods together to form a continuous model. We need to know how to get from one period to another. Both Keynes and Hicks resorted in one way or another to a link via expectations, though both provided what now appear to be inadequate explanations of their formation.

The modern new classical theory of macroeconomic equilibrium avoids the short and the long run distinction by appealing both to the perfectibility of markets per se, and to a theory of expectations which is itself based on perfect markets. Thus although the rational expectations approach 'solves' the problem of linking market periods, it does so in a way which avoids rather than solves the problems of market period analysis. Perfect

markets do not have dynamics with limited time horizons and rigidities and thus, in the rational expectations perfect foresight model, there is really no need for market period analysis. It is clear however that market imperfections in real variables and expectations do exist, and hence that short-run temporary equilibria cannot easily be linked together.

See Also

- ▶ Long Run and Short Run
- ▶ Marshall, Alfred (1842–1924)
- ▶ Reservation Price and Reservation Demand
- ▶ Temporary Equilibrium

Bibliography

Harrod, R. 1934. Doctrines of imperfect competition. *Quarterly Journal of Economics* 48: 442–470.

Hicks, J.R. 1956. Methods of dynamic analysis. In *25 Essays in honour of Erik Lindahl*. Stockholm. Repr. in J.R. Hicks. *Money, interest and wages*. Oxford: Basil Blackwell, 1982.

Hicks, J.R. 1965. The method of Marshall. In *Capital and growth*, ed. J.R. Hicks. Oxford: Clarendon Press.

Marshall, A. 1920. *Principles of economics*. 8th ed. London: Macmillan.

Market Power and Collusion in Laboratory Markets

Douglas D. Davis

Abstract

Despite the robust tendency of laboratory markets to generate competitive outcomes, some market designs deviate persistently from competitive predictions. This article discusses the primary drivers of supra-competitive prices that have been observed in market experiments.

JEL Classification

C9; L1

The robustness of competitive market predictions stands as one of the most impressive results in experimental economics. Laboratory markets regularly generate competitive outcomes in environments populated by just two or three sellers. However, as in natural contexts, competitive outcomes do not always emerge. This article reviews results of laboratory markets in which price increases are driven by factors such as the exercise of unilateral market power or by collusion.

Before reviewing the main concepts and contributions in this area, I offer two observations. First, laboratory methods represent an important but limited complement to existing empirical tools for investigating market performance. Given the stark simplicity and limited duration of laboratory markets, experimentalists can aspire to say little about specific naturally occurring markets. Experiments can, however, provide important insights into the behavioural relevance of theories upon which antitrust policies are based.

Second, the trading rules defining negotiations and contracting can exert first-order effects on market competitiveness. For example, markets organized under the double auction trading rules used in many financial exchanges, are much more robustly competitive than markets organized under the posted-offer trading rules used in most retail exchanges: duopoly or even monopoly sellers are less able to increase market prices in double-auction than in posted-offer markets (Davis and Holt 1993, chs 3, 4; Holt 1995). Indeed, one of the motivating factors in the emerging field of institutional design was an interest in developing institutional rules that promoted efficient market outcomes.

For specificity I focus here on results from posted-offer markets, primarily because posted-offer markets allow a particularly intuitive illustration of the factors affecting market competitiveness. However, a host of other trading institutions exist, ranging from single and multi-unit auctions, to multi-sided computerized 'smart' markets, and again to institutions that exist primarily as theoretical constructs, such as quantity-setting Cournot mechanisms. The competitive implications of each of these institutions must be evaluated independently.

Posted-Offer Markets and Unilateral Market Power

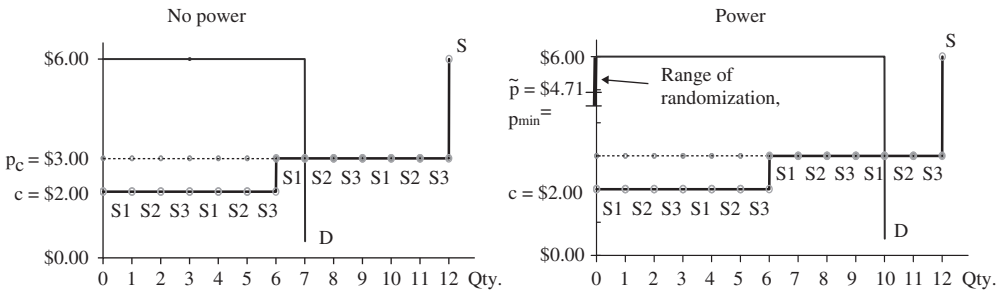
Unilateral market power is perhaps the most frequently observed reason why prices in laboratory markets deviate from competitive predictions. This market power exists when one or more sellers, acting on their own, find it profitable to raise prices above the competitive level. The supply and demand structures shown in the two panels of Fig. 1 illustrate how capacity restrictions can create market power. In each panel, the market consists of three sellers, S1, S2 and S3, each of whom offers four units for sale, under the conditions that two units cost \$2.00 and two units cost \$3.00. A buyer will purchase a fixed number of units (seven in the left panel or ten in the right panel) at prices less than or equal to \$6.00.

Exchange in these markets proceeds in a number of trading periods. At the outset of each period, sellers simultaneously make price decisions. Production is 'to order' in the sense that sellers incur costs only for the units that actually sell. Once all sellers post prices, a simulated fully revealing buyer makes all possible purchases, starting with the least expensive units first. In the case of a tie, the buyer rotates purchases among the tied sellers.

In the market shown in the left panel of Fig. 1 the buyer will purchase at most seven units. Given an aggregate supply of 12 units, sellers in this market have no market power: at any common price above \$3.00, each seller can increase sales from an expected 2.33 units to four units by posting a price just slightly below the common price. For any vector of heterogeneous prices above \$3.00 only the seller posting the lowest price will sell all four units. The seller posting the second highest price will sell three units, while the high-pricing seller will sell nothing. The unique Nash equilibrium for the stage game has each seller posting the competitive price of \$3.00, selling 2.33 units in expectation and earning \$2.00.

Expanding demand to ten units, as shown in the right panel of Fig. 1, limits excess supply, and thus creates market power. Given that the highest price seller is now certain to sell at least two units, the competitive price of \$3.00 is no longer a Nash equilibrium for the stage game. At a common price of \$3.00 each seller sells 3.33 units (in expectation) and earns \$2.00. By posting a price of \$6.00, any seller can sell two units and increase earnings to \$8.00. A common price of \$6.00 is not an equilibrium for the stage game, since any seller would find that deviating from \$6.00 increases sales to four units. Sellers have similar incentives to undercut any common price down to a minimum $p_{min} = \$4.50$, where the profits from selling four units as the lowest pricing seller equals earnings at the limit price. The equilibrium for this game involves mixing over the range from \$4.50 to \$6.00. As shown in the figure, the unique symmetric equilibrium is \$4.71.

An extensive series of experiments show that sellers respond to unilateral market power by raising prices. Further, power drives pricing outcomes more powerfully than do changes in the number of sellers. For example, when they reallocated units among five sellers to create market power, Davis and Holt (1994) observed substantial price



Market Power and Collusion in Laboratory Markets, Fig. 1 Supply and demand arrays for markets without and with unilateral market power

M

increases. However, reducing the number of sellers from five to three in a way that held market power conditions fixed, Davis and Holt observed only modest additional price increases. Market power of the sort illustrated in the right panel of Fig. 1 has wide applications, ranging from distortions in markets for emissions trading (Godby 2000) and for electricity transmission (Rassenti et al. 2003), to price stickiness in the face of aggregate demand shocks (Wilson 1998).

Tacit Collusion

Experimentalists have also observed supra-competitive prices in repeated market games where sellers have no market power. This *tacit collusion* has been observed most frequently in duopolies (for example, Alger 1987; Fouraker and Siegel 1963). However, tacit collusion has also been observed in thicker markets where sellers possess no market power. For example, Cason and Williams (1990) observe persistently high prices in a four-seller design similar to that shown as the left panel of Fig. 1. Experimentalists often measure tacit collusion as the difference between observed prices and prices consistent with the Nash equilibrium for the market analysed as a stage game. Importantly, other than exceeding equilibrium price predictions, tacitly collusive laboratory outcomes typically exhibit no obvious signs of coordinated activity.

Tacit collusion may coexist with market power. For example, prices in the market power sessions reported by Davis and Holt (1994) were significantly above prices consistent with the equilibrium mixing distribution. In this context, the difference between mean observed prices and the mean of the equilibrium mixing distribution may be reasonably taken as a measure of tacit collusion.

Tacit collusion is not yet well understood, and isolating the causes of tacit collusion represents an important project for future experimental work. Price signalling activity at least partially explains tacit collusion (for example, Durham et al. 2004). However, evidence suggests that more than price signals and responses may be at play. Dufwenberg and Gneezy (2000) report an experiment where duopolists deviate from the static Nash (competitive) prediction for a game, even when sellers are rematched into different markets after each decision. In such a context price signalling is not possible.

Explicit Collusion

Given opportunities to explicitly discuss pricing, laboratory sellers quite persistently organize profit-increasing cartels (Isaac et al. 1984). However, a capacity to monitor agreements and prevent secret discounts appears critical to the success of these arrangements (Davis and Holt 1998). Given the illegality of explicit agreements, the more interesting questions regarding explicit collusion concern the capacity of authorities to detect such arrangements through the actions of sellers in the market (Davis and Wilson 2002).

Other Factors Affecting Pricing

A host of experimental studies indicate that standard 'facilitating practices' can contribute to price increases. Experimental studies where supra-competitive prices have been attributed to facilitating practices include 'most favoured nation' and 'meet-or-release' clauses (Grether and Plott 1984), non-binding price signals (Holt and Davis 1990) and multi-market competition (Phillips and Mason 1991).

Buyer behaviour can also affect market outcomes. When buyer decisions are simulated, details of the purchasing rules can have a large effect on prices (Kruse 1993). Powerful human buyers can substantially undermine both market power and tacit collusion (Ruffle 2000). However, the use of real rather than simulated buyers appears to generate more competitive prices even when the human buyers engage in no strategic behaviour (Coursey et al. 1984).

Finally, information conditions and even sellers' expectations can significantly affect pricing outcomes. For example, Huck et al. (2000) report that information regarding underlying supply and demand conditions facilitates the exercise of predicted market power (markets are drawn to static

Nash predictions). However, information on rival sellers' profits made markets more competitive in a market where the high-profit seller has the highest market share, so imitation by others will tend to expand quantity and reduce price. Also, in a Cournot context, Huck et al. (2007) report that seller aspirations for increased profits helped consolidated sellers maintain prices substantially above static Nash levels.

See Also

- ▶ Anti-trust Enforcement
- ▶ Bertrand Competition
- ▶ Experimental Economics
- ▶ Market Institutions

Bibliography

Alger, D. 1987. Laboratory tests of equilibrium predictions with disequilibrium price data. *Review of Economic Studies* 54: 105–145.

Cason, T.N., and A.W. Williams. 1990. Competitive equilibrium convergence in posted-offer markets with extreme earnings inequities. *Journal of Economic Behavior and Organization* 14: 331–352.

Coursey, D., R.M. Isaac, M. Luke, and V.L. Smith. 1984. Market contestability in the presence of sunk (entry) costs. *RAND Journal of Economics* 15: 69–84.

Davis, D.D., and C.A. Holt. 1993. *Experimental economics*. Princeton: Princeton University Press.

Davis, D.D., and C.A. Holt. 1994. Market power and mergers in laboratory markets with posted prices. *RAND Journal of Economics* 25: 467–487.

Davis, D.D., and C.A. Holt. 1998. Conspiracies and secret discounts in laboratory markets. *Economic Journal* 108: 736–756.

Davis, D.D., and B. Wilson. 2002. An experimental investigation of methods for detecting collusion. *Economic Inquiry* 40: 213–230.

Dufwenberg, M., and U. Gneezy. 2000. Price competition and market concentration: An experimental study. *International Journal of Industrial Organization* 18: 7–22.

Durham, Y., K. McCabe, M.A. Olson, S. Rassenti, and V. Smith. 2004. Oligopoly competition in fixed cost environments. *International Journal of Industrial Organization* 22: 147–162.

Fouraker, L.E., and S. Siegel. 1963. *Bargaining behavior*. New York: McGraw-Hill.

Godby, R. 2000. Market power and emission trading: Theory and laboratory results. *Pacific Economic Review* 5: 349–364.

Grether, D.M., and C.R. Plott. 1984. The effects of market practices in oligopolistic markets: An experimental examination of the ethyl case. *Economic Inquiry* 24: 479–507.

Holt, C.A. 1995. Industrial organization: A survey of laboratory research. In *The handbook of industrial organization*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.

Holt, C.A., and D.D. Davis. 1990. The effects of non-binding price announcements on posted-offer markets. *Economics Letters* 34: 307–310.

Huck, S., H. Normann, and J. Oechssler. 2000. Does information about competitors' actions increase or decrease competition in experimental oligopoly markets? *International Journal of Industrial Organization* 18: 39–57.

Huck, S., K.A. Konrad, W. Mùller, and H.T. Normann. 2007. The merger paradox and why aspiration levels let it fail in the laboratory. *Economic Journal* 117: 1073–1095.

Isaac, R.M., V. Ramey, and A. Williams. 1984. The effects of market organization on conspiracies in restraint of trade. *Journal of Economic Behavior and Organization* 5: 191–222.

Kruse, J.B. 1993. Nash equilibrium and buyer rationing rules: Experimental evidence. *Economic Inquiry* 31: 631–666.

Phillips, O.R., and C.F. Mason. 1991. Mutual forbearance in experimental congolomerate markets. *RAND Journal of Economics* 23: 395–414.

Rassenti, S.J., V.L. Smith, and B.J. Wilson. 2003. Controlling market power and price spikes in electricity networks: Demand-side bidding. *Proceedings of the National Academy of Sciences* 100: 2998–3003.

Ruffle, B.J. 2000. Some factors affecting demand withholding in posted-offer markets. *Economic Theory* 16: 529–544.

Wilson, B.J. 1998. What collusion? Unilateral market power as a catalyst for countercyclical markups. *Experimental Economics* 1: 133–145.

Market Price

G. Vaggi

Keywords

Cantillon, R.; Competition; Market price; Natural price

JEL Classifications

D4

The market price, or market value, is defined as the actual price paid for a commodity during a certain period of time, and may be contrasted with

the natural or normal price, which is determined by the long-term forces and the permanent causes of the value of commodities (see ► Natural Price).

The distinction between market price and the intrinsic value of a good can be traced back to the origins of economic science. Before Adam Smith, Richard Cantillon had already analysed the causes which influence the temporary value of a commodity (Hollander 1973, p. 41).

Many different causes can affect the market price of a commodity and it is difficult to explain the day-to-day changes in its value. However, economic theory has generally singled out the relationship between the demand for a product and its supply on the market as the main force determining the market value. For Smith, the existence of a positive difference between the effectual demand for a commodity and the quantity of it which has been produced and brought to the market leads to a high market price vis-à-vis the natural value, and vice versa. But if there is free competition between producers, market prices cannot be too different from natural prices for a long period of time. Market competition forces lead to the gravitation of market prices around the natural prices. Therefore in classical political economy the two concepts are carefully listed. In particular, the market price is continuously brought towards the natural price.

The concept of market price is an important feature of Adam Smith's description of the competitive mechanism and the way in which it leads to a uniform rate of profit in all sectors of the economy. For instance, when the supply of a commodity falls short of its effectual demand the market price is higher than the natural one, because of the competition between the buyers who are eager to purchase that good (Smith 1776, pp. 73–4). Either one or more of the three component elements of value – wages, profits and rent – is paid at a rate higher than the natural rate. In a freely competitive economy producers compare their rate of profit with profit rates earned in other activities. Thus entrepreneurs invest their capital in the sectors which yield the highest rates or profit. This leads to an increase in the output of the commodities whose market prices are higher than natural ones, and vice versa a decrease when market prices are lower than natural values. Therefore the concept of market price is part of Smith's explanation of the changes in output which occur from one production period to another in each sector.

Given the natural price, and the corresponding level of effectual demand, the increase in output leads to a market situation in which more consumers (willing to pay for the good at its natural price) can be satisfied. There is less competition than before between the consumers, and the market price tends to move towards the natural price. Again, the entrepreneurs compare market prices and profit rates in all sectors of the economy and capital will move if the rate of profit is not uniform. Only when all the demand is matched by an equal supply at a market price level which equals the natural one will competition stop; in this situation the market price is exactly the right amount to pay all the components of price at their natural values. The market price depends on excess demand (or supply) on the market at any moment in time, but cannot be too far away from the natural price for a long period of time, because competition tends to bring it towards this level.

Alfred Marshall's distinction between the market and normal value of commodities is similar to Adam Smith's. Normal value is related to the cost production of commodities, while market prices are mainly influenced by utility and demand (Marshall 1920, pp. 289–90). Marshall also believes that it is difficult to work out a precise theory to determine the market values of commodities; they are affected by too many factors. Marshall argues that there are long and short period forces acting on prices. But he is much more sceptical than Smith about the existence of a precise mechanism, namely competition, which should prevent short-term market prices from moving too far away from the normal price of commodities.

See Also

► Natural Price

Bibliography

Hollander, S. 1973. *The economics of Adam Smith*. London: Heinemann.

Marshall, A. 1920. *Principles of economics,* 8th ed. Reprinted, London: Macmillan, 1972.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford University Press, 1976.

Market Share

William G. Shepherd

The leading element of market structure, the market share of the firm is a simple fact which is central to the study of industrial organization. Because it shows how far the firm's control over the market extends, market share is the direct indicator of each firm's position in the market. Its key role is of ancient tradition, and it reflects a universal recognition in business life that market share is commonly decisive for the firm's degree of success.

Market share's role derives from the fundamental theory of monopoly. A pure monopoly (market share of 100 per cent) controls the entire market, exerting the maximum monopoly power that is possible within the specific conditions of that market. The monopolist's demand is as inelastic as that of the entire market, and so its ability to raise price and influence other market outcomes reflects a full exploitation of that inelasticity.

The inelasticity arises from a lack of near substitutes, from instilled consumer loyalties, and from lags in adjustments. The inelasticity allows the monopolist to use price discrimination so as to maximize profits and to minimize the possibility of new competition from any other firm. The monopolist may engage in a variety of strategic actions, involving the full range of prices and various elements of product quality, marketing, and threats. Also, the monopolist is often able to raise barriers to new competition, by using strategic actions and adjusting the amount and technology of the capital that it installs. These sources of monopoly power are long-established in theory and thoroughly familiar in the mainstreams of actual industrial activity.

All of these factors also operate at lower market shares, though with lower degrees of force. Dominant firms (conventionally, those firms with at least 40 per cent of the market and no close rival) exercise a high degree of monopoly power, though less than pure monopolists. Their use of strategic actions to forestall small rivals or new entrants is often forceful and complex, but the effectiveness shades down directly in line with the dominant firm's market share.

Firms with market shares down in the 20 to 40 per cent range usually have significant market power, but they often face other substantial firms and problems of oligopolistic interdependence. Here market power is usually obtained mainly via collusion with other oligopolists. Below 20 per cent, and particularly below 10 per cent of the market, firms hold little or no monopoly power. This is generally true regardless of the market shares held by other firms.

Market share is therefore the most general, direct single indicator of the firm's ability to exert market power. It is mainly an ordinal indicator: within each market, each firm's market power varies with its market share. Cardinal comparisons of market power among markets do not operate along such a single scale. A 60 per cent share of one market may give much higher monopoly power than the same share in another market. Even so, market dominance usually provides a high degree of market power in every market.

Because market share is so crucial, it is a highly sensitive datum. Firms almost never willingly disclose their shares for fear of giving advantage of some sort to competitors or of inviting unwanted policy actions. Therefore, reliable market shares are scarce. The Census is prohibited from revealing them from the data that it collects. Some market shares emerge from commercial market surveys and from antitrust cases, as well as from occasional painstaking scholarly studies. But market-share data have been far scarcer than concentration ratios, and so those ratios took centre stage in research from the 1930s to the 1970s in the US. Market shares, and the role of individual firm dominance, therefore fell into relative neglect during this period, and oligopoly condtions drew disproportionate attention. Market-share research

has expanded since 1970, but it usually has to rely on rather rough estimates of market shares. Such research clearly has much further to go, in improving data and exploring the causes and effects of market shares.

Even so, the research has already succeeded in affirming market share as the central element in market structure. This has been estimated with regression models including the main structural elements. The most general forms are two. One is

$$\text{Profit Rate} = a + b\text{MS} + c\text{Conc} + d\text{Size} + e\text{AdInt} + f\text{Growth} \quad (1)$$

(where Profit Rate is the rate of return on equity capital or total assets, MS is market share, Conc is 4-firm concentration in the market, Size is a measure of the firm's absolute size, AdInt represents the firm's advertising as a percentage of its sales, and Growth is a filter variable for the firm's growth rate). The other form is

$$\text{Profit Rate} = a + b\text{MS} + c\text{Conc} + d\text{HB} + e\text{MB} + f\text{Growth} \quad (2)$$

(where HB and MB are dummy variables for high and medium entry barriers). In both forms, the *d* and *e* variables represent entry-impeding conditions. These equations are fitted to cross-section data for 100 or more large industrial firms, across a range of industries. The tests also pool up to 10 years' data for various periods since 1960, in order to reflect basic conditions rather than yearly fluctuations.

Various studies report consistent findings. Market share's partial correlation with profit rates is very highly significant, with *t*-ratios commonly above 9.0. The coefficient is usually in the 0.2 to 0.3 range, showing that an added 10 points of market share commonly yield profit rates two to three points higher. The a term represents the competitive rate of return, at the prevailing cost of capital. The analysis therefore shows that a small-share firm might earn a 10 per cent profit rate, while a 60 per cent market-share dominant firm might expect a 20 to 25 per cent profit rate. Such large differences accord with common business experience. The relationship appears to be linear, with market share's yield remaining steady virtually throughout the range of market shares. Tests with Tobin's *Q* ratios in place of profit rates give closely similar results: market share's role is strong.

The other elements (concentration and entry barriers) play much weaker roles, in the tests of general conditions. Market share supplants concentration, on the whole, while entry conditions only affect profit rates by a point or two.

All of these results have been tested for the effects of risk, leverage, and other side factors. Given the amount of error in the profit and structural data, and the variety of company groups, time periods, and the alternative measures that have been tested, the research findings can be regarded as quite strong and consistent.

Two other tests are of interest. Market shares have been related to innovative activity, mainly as shown by patents and by new products. A curved pattern has emerged, with innovative activity peaking at market shares in the vicinity of 20–25 per cent. The influence causing changes in market shares have also been explored. Broadly, high market shares undergo a general process of erosion, usually in the range of one point per year. Dominance therefore dwindles, but rather slowly. High profit rates also encourage a slightly faster decline, as the firm cashes in on its market share rather than retain it by restraining its profitability.

While market share's role has emerged as central, the meaning of that for efficiency has become highly controversial. A high share may reflect a mere seizing of control, by means of mergers, anticompetitive acts, or sheer luck. The resulting monopoly effects may be harmful and without redeeming causes. That is the conventional view. Against it, the neo-Chicago School claim is that dominant firms arise because of economies of scale or some superiority (management, innovations, etc.) which gives them supremacy in the market. The dominance reflects efficient causes, which justify whatever monopoly effects may result.

Both views admit that the origins may include both monopolizing and efficient causes, while various monopoly effects may occur. The clash is over the amounts of these causes and effects. Mainstream experts rate the efficient causes low and the monopoly effects high, while neo-Chicagoans see all causes as efficient and all monopoly effects as trivially small. Chicagoans have provided little empirical basis for their claim, and so the burden of proof is still against them. But the issue is open.

The measurement of market shares begins with the effort to define the relevant market. A market is a group of buyers and sellers exchanging goods which are highly substitutable, perfectly so in the ideal case. The market's edges are set by the zone of choice of the main mass of buyers, as they compare alternative products and suppliers.

Each market exists in two main dimensions: product space and geographical area. One first identifies the products which are highly substitutable and distinct from others, so as to form a distinct market. One also determines the geographic area within which the main mass of customers can choose.

The cross-elasticity of demand is the basic concept for such decisions, in both dimensions. But is is rarely available, and so usually one must turn to other practical signs of substitutability, such as product features, pricing patterns, participants' views, transport costs, and actual shipping patterns. One may also make estimates, by *gedankenexperiment*, of the responsiveness of goods to significant price changes, during reasonable time intervals, as was claimed to be the method used by the US antitrust agencies after 1982. But the definition of markets, which in turn determines the resulting estimated market shares, remains unavoidably a matter of judgement in some degree. No formula or measurement device has yet proven satisfactory in itself.

Actual distributions of market shares in real markets tend to display a strong gradation of market shares, from top to bottom. Usually there is no distinct 'oligopoly group', of a few leading firms with roughly equal shares. Rather, the array of firms usually tapers down from the largest firm to a fringe of small ones. The degree of this asymmetry varies, but in the average case each of the largest handful of firms has about twice the share of the next largest. This has led some analysts to speak of a 'law' of market structure which embodies such a half-share gradation.

There is no official source of individual-firm market share data. Market survey firms do make private estimates for sale, and some of these figures do filter out in stories in the business press. The 'PIMS' data bank contains estimates by many firms about their own products' shares, and the Federal Trade Commission's Line of Business data for 1974–7 include market shares. Yet access to both of these data sets is tightly restricted. This author's estimate's for certain US manufacturing firms in 1961, 1968, 1972 and 1980 are available but approximate.

Despite all the secrecy and measurement problems, the leading cases of high market shares have been known for many decades, especially those in the US economy. From Standard Oil, American Tobacco, and US Steel as of 1910–20, down to ALCOA and United Shoe Machinery in the 1940s, and IBM, Eastman Kodak, General Motors, Procter & Gamble, Gillette, Campbell Soup and many others in the 1960s to the 1980s, the more notable dominant positions have become reliably known through antitrust cases and general trade information.

A relatively few prominent instances have somehow avoided the usual erosion of high market shares, by fair means, or foul, or both. They present a leading problem, both for research and for antitrust consideration. Market shares re-emerged in the 1970s as a leading research frontier in the field of industrial organization. They are likely to continue to be hard to measure, highly sensitive in policy matters, and intensely debated by scholars.

See Also

- ▶ Concentration Ratios
- ▶ Herfindahl Index
- ▶ Market Structure

Bibliography

Bain, J.S. 1956. *Barriers to new competition*. Cambridge: Harvard University Press.
Caves, R.E., and M.E. Proter. 1976. Barriers to exit. In *Essays on industrial organization in honor of Joe S. Bain*, ed. R.T. Masson and P.D. Qualls. Cambridge: Ballinger.
Caves, R.E., M. Fortunato, and P. Ghemawat. 1984. The decline of dominant firms, 1905–1929. *Q J Econ* 99: 523–546.
Gale, B.T., and B.S. Branch. 1982. Concentration and market share: Which determines profits and why does it matter? *Antitrust Bulletin* 27: 83–106.
Gaskins, D. 1971. Dynamic limit pricing: Optimal pricing under threat of entry. *Journal of Economic Theory* 3: 306–322.
Goldschmid, H.J., H.M. Mann, and J.F. Weston (eds.). 1974. *Industrial concentration: The new learning*. Boston: Little, Brown.
Kwoka, J.E. 1979. The effect of market share distribution on industry performance. *Review of Economic and Statistics* 61: 101–109.
Martin, S. 1983. *Market, firm and economic performance*, Monograph series in Economics and Finance. New York: New York University Press.
Mueller, D.C. 1983. *The determinants of persistent profits*. Washington, DC: Federal Trade Commission.
Peletzman, S. 1977. The gains and losses from industrial concentration. *Journal of Law and Economics* 20: 229–263.
Scherer, F.M. 1980. *Industrial market structure and economic performance*, 2nd ed. Boston: Houghton Mifflin.
Shepherd, W.G. 1972. The elements of market structure. *Review of Economics and Statistics* 54: 12–25.
Shepherd, W.G. 1975. *The treatment of market power*. New York: Columbia University Press.
Shepherd, W.G. 1979. *The economics of industrial organization*. Englewood Cliffs: Prentice-Hall.
US Department of Justice, Antitrust Division. 1982. *Merger guidelines*. Washington, DC, June 14.
US Federal Trade Commission. 1974–1977. *Line of business data*. Washington, DC: Annual.

Market Socialism

W. Brus

Market socialism is a theoretical concept (model) of an economic system in which the means of production (capital) are publicly or collectively owned, and the allocation of resources follows the rules of the market (product-, labour-, capital-markets). With regard to existing socialist economies the term is often applied more loosely to cover both systems which tend to approximate it in the strict sense (as the Yugoslav system in the aftermath of the 1965 reform), as well as those which replace commands and physical distribution of producer goods by financial controls and incentives as instruments of central planning ('regulated market', as in the Hungarian 'new economic mechanism' after the 1968 reform).

Introduction

Marx's political economy had for a long time been interpreted to hold that socialism is incompatible with the market. Market relations, even in their simplest form of commodity exchange between two self-employed producers, are presented in *Das Kapital* (Marx 1867) as a nucleus out of which – logically and historically – capitalism emerges. The market forms an indispensable link between economic actors when they are apparently separated from each other (by private ownership), and the social nature of their activity is hidden, revealing itself and being verified only through exchange; the overall outcome is then an ex post resultant of a multitude of spontaneous actions, with the negative consequences becoming the more pronounced the more developed the truly social character (interdependence) of the economic process. Socialism – according to this line of thought – makes the market redundant and overcomes its shortcomings as an allocation mechanism by bringing into the open the social nature of work, assigning it directly ex ante to a particular role in the economic process through the 'visible hand' of planning, which secures full utilization of resources, free of cyclical fluctuations. Above all, socialism removes the absurdity of having, side by side, unsatisfied needs, and excess capital and labour.

After the Russian revolution of 1917, when the shape of socialist economic systems became a practical problem, basic elements of the marketless concept of socialism found their way into programmatic documents of the Communist

parties. Any application of the market mechanism was presented as only a temporary concession, to be justified mainly by the immaturity of the socio-economic conditions that required a longer transition period between capitalism and socialism, especially in underdeveloped countries with dominant peasant agriculture and other types of 'petty commodity production' (Communist International 1929). At the same time, however, the social-democratic wing of Marxism began to recognize the relevance of the market for the operation of a socialist economy (Kautsky 1922).

Theoretical debates on market socialism acquired a new dimension in the interwar period, particularly after republication by F.A. von Hayek (1935) of an article by L. von Mises published originally in 1920, which categorically denied the possibility of rational economic calculation under socialism, because exchange relations between production goods and hence their prices could be established only on the basis of private ownership. Among the many attempts at refutation of this view (Taylor 1929; Dickinson 1933; Landauer 1931; Heimann 1932), probably the best known is that by Oskar Lange (1936–7). Similar ideas had been developed in the same period by Abba Lerner (1934–7), hence the often used designation of 'Lange–Lerner solution'.

Lange not only denied the purely theoretical validity of von Mises' stand (by pointing to Barone's (1908) demonstration of the possibility of dealing with the question through a system of simultaneous equations), but also tried to present a positive solution. This was to consist of a 'trial and error' procedure, in which the Central Planning Board performs the functions of the market where there is no market in the institutional sense of the word (or – it may be added – where market imperfections threaten the parametric function of prices). In this capacity the Board fixes prices, as well as wages and interest rates, so as to balance supply and demand (by appropriate changes in case of disequilibrium), and instructs managers of socialist enterprises (and entire industries) to follow two rules: (1) to minimize average cost of production by using a combination of factors which would equalize marginal productivity of their money unit-worth; (2) to determine the scale of output at a point of equalization of marginal cost and the price set by the Board.

The emphasis in the elaboration of the 'trial and error' procedure was, in the first place, on proving socialism to be capable of allocating resources in a way equivalent to a purely competitive market system. But it acquired much wider significance as an attempt to construct a normative model of market socialism. However, in the latter interpretation – as a normative solution – the model of a socialist market economy becomes more vulnerable to practical tests of validity than would a model of a capitalist market economy. The real behaviour of actors in capitalist markets is by no means determined by the propositions of general equilibrium theory, whereas socialist managers are to be *instructed* to follow the textbook rules, with all the ensuing consequences of iteration processes which may not only operate with considerable time-lags and oscillations, but may not be convergent at all. This explains to some extent the inclusion into the model of a number of features which would distinguish it from a standard 'free market'. In Lange's presentation the main such feature is the determination of the rate of accumulation, not by market processes but directly ('arbitrarily') by the Board, which establishes also the rules of distribution of the social dividend from publicly owned capital and land (with a proviso that this should not affect the choice of occupation).

In addition, the 'trial and error' procedure assumes that actual markets are limited to consumer goods and labour only, while the functions of the market for production goods are performed by the Board itself (market *simulation*). Moreover, in a generalized version, even these assumptions may be dropped. The Board may impose its own scale of preferences on the composition of consumer goods output, even ration goods and assign people to their jobs, but it can still apply the 'trial and error' procedure to derive accounting (shadow) prices of production goods provided that there is no rationing outside the sphere of distribution of consumer goods and labour, i.e. at least a simulated market for production goods must exist. Thus, the market may be understood in Lange's model as a 'computing device of the

pre-electronic age' for solving a system of simultaneous equations, as that author himself emphasized in his last article (Lange 1965), in which the relative merits of the computer and the market are weighed very carefully, with the market in an institutional sense being judged by no means superior on all scores, particularly with regard to long-term dynamic problems.

Market socialism in the above form was to be capable of combining the allocative efficiency of competitive conditions (secured by the Board's rules, which ought to preclude oligopolistic and monopolistic behaviour), with welfare maximizing income distribution (by eliminating inequalities stemming from private ownership of capital and land) and internalization of externalities (by inclusion of all alternatives foregone into comprehensive social cost calculations). An economy operating on the principles of this model was to be open to innovations without generating cyclical fluctuations. The main difficulty considered – the danger of bureaucratization of economic life – was assessed (by Lange) as not greater than that under monopolistic capitalism, and perhaps even more containable under socialism due to democratic control over public functionaries.

This concept of market socialism came under heavy criticism from two opposite sides: from those who disputed the validity of the socialist component in the market system, and those who disputed the market component in the socialist system. The first kind of criticism, mainly following Hayek (1940) has concentrated on the unlikelihood of creating the informational and motivational foundations of market-type managerial behaviour without the background of private ownership which provides the necessary stimuli (expected returns) and constraints (financial responsibility) to innovative decisions involving risk. Schumpeter (1942) denied the relevance of the charge by pointing to the divorce between ownership and management under modern capitalism, but the empirical evidence from communist countries suggests that this is indeed a most serious issue. The second type of criticism, apart from that of general ideological nature, has concentrated on the market-type behaviour postulated in the model and directed toward static efficiency, and the overriding exigency of a dynamic process with full utilization of resources which it is claimed can be satisfied only through direct central planning – otherwise, strong elements of instability would become inherent in the system, along with deviations from the postulated pattern of income distribution (Dobb 1939; Baran 1952). Independently, the rationale of relying on market mechanisms of allocation in the face of large-scale dynamic problems was widely questioned in the Soviet debates in the 1920s (Erlich 1960), as well as in the East European countries and in China after World War II. Lange himself acknowledged the need to re-examine his model from the point of view of long-term economic dynamics, oscillation, and income effects (Lange 1947 and 1965). As for the informational and motivational weaknesses stemming from elimination of private ownership, the problems involved had not attracted much attention in communist countries until the last quarter of the 20th century, when they surfaced quite distinctly in connection with difficulties encountered in the course of various attempts to reform the economic system by increasing the role of actual markets.

The Command System

The history of economic institutions in communist countries could be interpreted as displaying a certain tendency toward broadening the scope of operation of the market; but changes have been slow, and by the mid-1980s most communist countries still adhered to the essentials of the orthodox Soviet system based on commands and physical allocation of producer goods. The command system introduced in the USSR in the early 1930s was transplanted after World War II to all other communist countries (including China and Cuba), which might be taken as evidence that it was regarded as a general model and not as a reflection of Russian conditions peculiar to that time. Prior to that the Soviet Union went through so-called 'war communism' (during the civil war, 1918–20), when circumstances of extreme penury precipitated an attempt to switch to a moneyless economy with resources and products distributed

in natura (it was this system which first prompted von Mises to challenge socialism's capacity for rational economic calculation). Later came the period of 'the new economic policy' (NEP), (1921–28/9), when the market was allowed to function relatively widely, but only as a temporary expedient of transition from capitalism to socialism. The first five-year plan (1928–32) and forced collectivization of agriculture marked the end of this period; the command economic system was installed as corresponding to the stage of socialism (a lower stage of communism).

The principles of the command system as a model do not eliminate the market completely, but they relegate it to the peripheries of the state-owned (or fully state-controlled, as in the case of nominally collective farms and other cooperatives) production sphere. Freedom of choice of consumer goods – outside public consumption – combined with freedom of choice of occupation and jobs means that economic relations between the state and the households have to go through some kind of market with an active role for money. Prices, wages and interest rates on savings and personal loans etc. affect choices made by households as labour suppliers, consumers and savers. Free sales of agricultural produce above the state-quotas, particularly from individual plots cultivated by members of collective farms (as well as by many state employees) constitute another important segment of the market in the command system. In practice, however, all these segments are subject to restrictions, such rationing, forced labour, curbs on labour mobility, constraints on the scale of individual farming etc. During the Stalinist period these restrictions were very severe, at some points overshadowing the elements of the market. However, the general tendency since 1953 has shown a gradual removal of restrictions, which makes the model's assumptions more meaningful.

Within the production sphere of the command system, the market does not function as an allocative mechanism. This role belongs to the Plan, which is meant to decide in a direct way not only major macroeconomic issues of growth, structure of productive capacity and income distribution, but also detailed schedules of current output and input, together with directions of flows between production units and toward the consumption sphere – predominantly in physical terms. Plans in the command system are in fact commands. The supply of production factors and intermediate goods is limited by rationing (allocation orders), and performance is assessed by plan-fulfilment yardsticks. Money is used within the production sphere for aggregation and accounts-control purposes, and the forms of exchange transactions (sales, purchases, prices, credit) between enterprises are used in the same way. However, money remains *passive,* i.e. calculations and transactions in money terms follow the physical flow of resources and intermediate goods decided by the Plan (Brus 1961); this means also that although among the targets and limits of the Plan financial goals (costs, profit, etc.) figure as well, the latter are subordinated to the physical indicators, and the financial position of enterprises is adjusted (through subsidies, dual price systems, differentiated product taxes, etc.) in such a way as to enable them to fulfil the physical tasks – the 'soft budget constraint' (Kornai 1980). Thus, the production sphere is separated from the market elements outside it (relations between the state and the households, and between the households), and consumers' choices are not transmitted to the producers via the market mechanism, but filtered through planners' preferences. Similarly, the domestic economy is separated from foreign markets, both Western and intra-Comecon, by the 'monopoly of foreign trade' which operates through import and export quotas and adjusts the financial position of importers and exporters via a 'price equalization mechanism' (Wiles 1969).

Consequently the command system fails to provide even the minimum conditions for Lange's 'trial and error' procedure: it can accommodate some kind of market in the consumption sphere with possibility of finding equilibrium prices for consumer goods, but it eliminates the market within the production sphere, where goods are rationed and prices are arbitrary. Obviously, the separation between the two spheres can never be complete; the feedback effect is particularly noticeable via wages which have – under normal circumstances and with all reservations due to

imperfection of labour markets anywhere – to reflect supply and demand, while at the same time constituting the major component of cost calculations that enter in one way or another into the considerations of the planners.

The Yugoslav Experiment

Market socialism first appeared as a practical challenge to the command system in the early 1950s in Yugoslavia, after the Stalin–Tito break. The primary motives of this challenge were not economic, although the economic difficulties arising from originally the most complete (for Eastern Europe) and – paradoxically – voluntary transplantation of the command system to a small country without resources on the Soviet scale, played a considerable part. The Yugoslav Communist Party searched mainly for political and ideological self-determination vis-à-vis the hitherto unquestioned authority of Stalin in the communist world. It was found in the concept of self-management, presented as an embodiment of this strain in Marxian ideas which emphasizes socialism as social order which overcomes alienation of labour by placing means of production under control of 'associated direct producers' (Ljubljana Programme 1958).

Contrary to Soviet doctrine, nationalization came to be regarded here as only the first step towards the socialization of the means of production, because even a socialist state is merely an indirect representative of the producers who remain wage labourers until they themselves decide how to use the means of production entrusted to them, and how to allocate the income generated. The process of socialization is thus tantamount to consistent development of self-management in every unit of the economy (and in other spheres of social life); the direct economic involvement of the state has to be curtailed gradually through decentralization of decision-making, not only with regard to current operations of enterprises, but also with regard to capital investment. The functions of the national plan are in principle only indicative, confined basically to provide information and framework for (voluntary) coordination, and to counteract monopolistic and oligopolistic behaviour; direct allocation of resources by state organs is an exception, for cases such as development aid to particularly backward regions or emergency measures in acute social situations. Otherwise the economy is to be regulated by the market, a *socialist* market – its participants being not private (individual or corporate) employers of labour, but associated producers, workers' collectives.

The process of implementation of these ideas was gradual and by no means straightforward. The problems of de-controlling prices and foreign economic relations, both essential for creating competitive conditions, proved to be particularly difficult. Despite numerous retreats in the field of prices, and reimpositions of controls on foreign operations, production in Yugoslavia ceased to be regulated by commands and input rationing, and money assumed an active role with prices tending to clear the market. Isolation from the outside world diminished substantially. In 1965 the country was launched into what was supposed to become the decisive stage of development of self-management and market socialism: the responsibility for 'expanded reproduction' (i.e. for the main bulk of capital investment) was to be shifted from the state budget to the self-managed units, which were to be free to decide about the shares of retained and distributed (as personal incomes) earnings, and about the use of the retained part. The mechanism of financial intermediation in the process of re-allocation of investment funds between sectors and areas was to be provided mainly by the network of commercial banks, with only a marginal role for the state budgets at various levels.

Yugoslavian market socialism aroused considerable interest and gave fresh impetus to theoretical debates, for example, to confrontations of this concept with the 'Lange–Lerner solution' (Bergson 1967). The behaviour of Yugoslav-type labour-managed firms was equated with (or held sufficiently similar to) cooperatives maximizing net income per member. Using assumptions of perfect competition, several authors beginning with Ward (1958) argued that a labour-managed firm pursuing its objective function will tend to

settle for a lower level of output and employment, and higher capital intensity, than would a capitalist firm in analogous conditions, and that it will even display a 'perverse' price-elasticity of supply (diminishing output and employment when the price of the product rises, and vice versa). Most of these peculiarities disappear however when imperfections of the market are taken into account. The labour-managed firm will try then to establish a maximum price (depending on the conditions of entry), and vary its output according to the movement of demand at that price, that is, in a 'normal' way (Lydall 1984).

Nevertheless, empirical evidence suggests that the attempt to combine market mechanism with self-management of the Yugoslavian kind generates problems unknown either to capitalist market economy or to full-fledged cooperatives operating in a market environment. They stem mostly from the fact that the workers' share in the enterprises' results is not based on any form of personal property rights, which they may carry with them, but exclusively on employment; upon termination of employment their stake disappears. This affects attitudes toward the distribution of returns between current personal incomes (for consumption or private savings), and collective investment; in particular, older workers and those without prospects or willingness to stay on the job will have a low propensity to invest out of the enterprise's income. The self-management organization of the economy also presents problems with regard to investment in other existing enterprises (there can be no sharing in profits), or in establishing new firms, especially in other sectors and regions (such firms become, as a rule, independent self-managing units). Absence of capital markets in which firms might participate directly puts even greater pressure on the banking system as a substitute. In the post-1965 period the banks were expected to establish themselves as fully fledged financial intermediaries, but the actual position proved rather disappointing, for reasons related at least to some extent to the specific features of the Yugoslav political system.

The one-party state used its power to impose a variety of formal and informal controls, for example through the so-called self-management compacts. Decentralization of state functions substantially enhanced the power of local organizations (particularly at the level of national republics and autonomous regions) which led to strong autarkic tendencies that not only had a disruptive effect on the unity of the national market, but also made it easier to overrule the commercial principles of operation (e.g. of the banks) by politico-administrative interference. This adversely affected the conditions of competition, especially as perennial balance of payments constraints frustrated the hopes of bringing competitive pressure from outside. At the same time, difficulties in promoting active participation in self-management of the workforces of large organizations gave rise to the so-called 'basic organizations of associated labour' (BOAL) – autonomous decisional and accounting units which may correspond to entire small or medium enterprises but form only self-contained parts of a large one. Excessive fragementation resulted in some cases, especially as links between workers' income and performance on a BOAL scale led to differences in remuneration for the same kind of work. In general, incentives linked to performance, particularly under imperfect markets, engendered problems that had been largely overlooked in Lange's model. This had assumed not only the viability of simulating perfect competition but also the existence of a motivational structure capable of inducing economic actors to observe fully the Board's rules, without any individual material stimulation beyond compensation of disutilities (the implicit interest in increasing the social dividend belongs to a different category of incentives). The scale and direction of change in income differentials – inter-enterprise, inter-sectorial, inter-regional – became a major issue not only in the Yugoslavian case but also in overall analysis of market socialism in comparison with both contemporary capitalist market economies and command systems.

For a considerable time Yugoslavian market socialism proved capable of combining fast growth with significant welfare gains that were unmarred by the shortages and glaring maladjustments so characteristic of command systems. However, the end of the 1970s and the beginning of the 1980s brought substantial

deterioration in this respect (slowdown of growth, high unemployment, accelerated inflation, fall in real earnings), which prompted renewed scrutiny of the effectiveness of the Yugoslavian model. In Yugoslavia itself the principle of selfmanagement was not subjected to open debate, although the question of property rights was raised again (Bajt 1982), and the role of political factors was quite widely recognized. With regard to the plan-market relationship, the predominant view seemed to be that the market had actually not been given a true chance, but accusations that excessive 'marketization' had precluded effective macroeconomic planning were also made (Mihailovic 1982).

The Hungarian Reforms

From the mid-1950s pressure to extend the role of the market began to manifest itself in countries belonging to the Council for Mutual Economic Assistance (CMEA), including the Soviet Union; towards the end of the 1970s a similar tendency appeared strongly in China. The reasons were basically economic – dissatisfaction with the performance of the economy under command systems, although in several cases (Hungary before the 1956 revolution, Poland in 1956–7 and again in 1980–81, Czechoslovakia in 1968) the presumed linkages between marketization of the economy and pluralization of the polity played an important part. Economic reforms – as the blueprints of the attempted changes came to be called – failed in most of the CMEA countries, or were reduced to rather secondary modifications within the framework of the command system. The failure was usually explained in academic literature by political resistance of the ruling elites, vested interests of the administrative state – and party – apparatus, coupled with reluctance on the part of the rank-and-file and managers to trade-off security for stronger incentives linked to efficiency. Difficulties of substance in devising and implementing a sufficiently consistent reform were mentioned less frequently, but they were certainly important and interacted with all other factors. By the mid-1980s among the CMEA countries only Hungary, where the 'new economic mechanism' (NEM) was introduced in 1968, could be regarded as actually outside the confines of command systems – despite the fact that the idea of market-oriented economic reforms kept returning in one form or another in most of the countries of the Soviet bloc, especially in response to crises. In 1981, during the existence of the independent trade union Solidarity, a wide ranging design of self-managed market socialism was worked out in Poland. A much more circumscribed reform, introduced after the suppression of Solidarity, met with a number of difficulties of both economic and political nature. China, having successfully revived the market mechanism in agriculture, embarked in 1984 upon a major programme of economic reform in industry and trade.

Conceptually, the Hungarian 'new economic mechanism' of 1968 is distinct from the Yugoslav market socialism, not only in leaving out self-management, but also in having a different relationship between the plan and the market. The principle of central planning is upheld, while the methods of realization are changed, with the market assigned an active role not only in relations between the state and the households (where the restrictions appearing in the practice of command systems are consistently removed), but also within the state production sector itself. Obligatory output targets for enterprises are abolished, as are physical allocations of production goods from the centre. Thus, enterprises are freed from hierarchical administrative commands and exposed to market-type self-regulatory mechanisms in their current operations, with profit as the main criterion and source both of incentives for the workforce (wage rises dependent on financial viability plus profit-sharing fund for bonuses) and of self-finance for autonomous investment. Prices, both in the consumption and production spheres, are meant to clear the market; but only some prices are allowed to fluctuate freely, and the most 'important' are fixed by state bodies, with other prices moving only within an established range. Isolation of internal from external markets is also lifted, again with substantial indirect controls retained. Thus, the question of incentives apart, the 'new economic mechanism' as a model meets

the Lange–Lerner requirement for the 'trial and error' procedure of establishing prices of production goods. Where it falls short of the Lange–Lerner solution is in the investment sphere: not only the rate of accumulation, but also the allocation of the *main bulk* of investment funds among sectors, areas and large individual projects is determined directly by the Board, whereas equilibrating supply of and demand for capital through appropriate variation of the interest rate takes a secondary place (only with regard to crediting enterprises' autonomous investment, which is considered secondary).

The capacity of the Board to harmonize economic activity on a micro-level with the general provisions of the plan is therefore supposed to rest on: (i) the macroeconomic framework created by the Board's fundamental decisions concerning distribution of national income (including principles of remuneration) and investment allocation; (ii) determination of 'rules of behaviour' for enterprises (success criteria and their incentive consequences) in such a way as to direct local interests onto a path convergent to general interests; (iii) fiscal, monetary and price policies which would effectively support (i) and (ii), in the first place by securing the parametric character of the 'indices of available alternatives' (Lange 1936–7), viz: prices, wages, interest and tax rates. The primacy of the plan so conceived means not only abandonment of direct forms of control, but elimination of central control as such over many aspects of economic activity (recognition of broad 'zones of indifference' as far as planners' preferences are concerned). The interaction between an effective central plan and a market mechanism which requires enterprises to adjust to general rules and conditions makes the model of *central planning with regulated market mechanism* (Brus 1961) an approximately adequate description of the concept of the 'new economic mechanism'.

These economic reforms were introduced in Hungary under mixed political circumstances. On the one hand, the party leadership became firmly committed to them, although the opposition was strong enough to force partial retreat from the principles of the reforms in the period 1973–8. On the other hand, the Soviet bloc offensive against the Czechoslovakian reforms of 1968 was an important adverse factor, among other things because it contributed to the abandonment of economic reforms elsewhere, leaving Hungary an exception within CMEA. The operation of the new economic mechanism was clearly affected by this, as Hungary had to adjust accordingly the management of her relations with other member-countries (particularly with the USSR), and with the CMEA as a whole. All this diminished the capacity of the new mechanism to respond to the deteriorating external conditions caused by the oil shocks of the 1970s and the Western recession. Hungary, poorly endowed in fuel and raw materials and at the same time highly dependent on foreign trade, was the worst hit country in Eastern Europe by the fall in the terms of trade, and the growing difficulties of exporting to the West.

Under the circumstances the performance of the Hungarian economy in the 1970s could be judged as relatively favourable, particularly in maintaining equilibrium on the domestic market. This was due in the first place to the successful development of genuinely cooperative activity combined with private initiative in agriculture, where the provisions of the reforms turned to the greatest advantage. With a rather broad consensus of opinion both among the political leaders and professional economists, the inconsistencies and retreats in implementation of the new economic mechanism began to be corrected at the end of the 1970s.

The most pertinent question however was whether, or to what extent, the failure of the systemic changes to live up to expectations was due to deviations from the 1968 blueprint, or to deficiencies in the blueprint itself. Special significance was attached to the search for reasons why, instead of applying the general rules and rigours of the market to state enterprises, the widespread practice was to tailor financial norms in such a way as to keep every enterprise afloat (cooperative enterprises were treated differently). This phenomenon, which replaced the former bargaining with the higher authorities over output targets and input allocations by new forms of bargaining over financial conditions, was noticed in the early stages of the reforms and attributed to the ideologically

motivated microeconomic job-security commitment (Granick 1975). However, the Hungarian debates at the beginning of the 1980s linked this also with the limitations on the investment activity of enterprises that was imposed by the principle of earmarking the main bulk of investment decisions for the central planner. An enterprise unsuccessful in its given line of business has only a very limited prospect on its own for restructuring or branching out if substantial capital outlays are involved, and this often narrows down the range of options to the stark choice between complete closure and subsidization, the latter course being that almost invariably taken.

Apart from the obvious softening of the 'budget constraint', with all ensuing consequences for the maintenance of pressure for efficiency in enterprises, and for distortions in market relations, this increased the enterprises' dependence on their administrative supervisors. In conjunction with criticism of the poor quality of many investment decisions taken by the centre (particularly when genuine *political* control is missing for lack of pluralism), this line of analysis convinced a substantial body of opinion in Hungary of the necessity to go beyond the product market (and the labour market in its existing form) to the creation of a *capital market.* Suggestions considered in Hungary in the first half of the 1980s envisaged a gradual and cautious movement along these lines, with a substantial part of investment ('infrastructural') still in the hands of the centre, and careful control over institutions of financial intermediation (commercially acting banks in the first place, but also direct issuance of bonds, and even equity shares in prospect). Nevertheless, the debates pointed clearly in the direction away from the mixed model of central planning cum regulated market mechanism, towards full-fledged market socialism, in which allocation of capital is accomplished through market instruments, with the rate of interest equilibrating supply and demand, as in Lange–Lerner. In similar vein, the *labour market* should provide the means to arrive at the equilibrium level of wages through the process of bargaining between management and the workforce; the latter – lacking the countervailing power of independent trade unions – would be able to make use of widely opened job-opportunities outside the state sector ('second economy') as a market instrument of pressure.

Conclusions

Thus, the evolution of both the Yugoslav self-management system and of the Hungarian economic reforms brings back on the agenda most of the problems debated theoretically in connection with the Lange–Lerner model of market socialism. Assuming that institutionalization of the capital market proves feasible in the framework of predominantly public ownership, the old question arises again of whether such a market, even with the help of fiscal and monetary tools of state intervention, is capable of securing continuously a macroeconomic level of demand appropriate for sustained economic growth with full employment, a goal that is regarded as an essential feature of socialism. Moreover, as any realistic concept of market socialism has to include incentives that are in some way linked to performance, capital markets and labour markets of the type referred to above must strongly affect the pattern of income distribution and of wealth as well. However, the assumption of the compatibility of capital markets with public ownership cannot be taken for granted, not only in view of the theoretical reasons advanced in the past, but to a considerable degree in the light of the practical experience of communist countries, where few instances of the relative success of fledgling capital markets can be found exclusively outside the state sector, whereas attempts to use them within that sector (e.g. the Yugoslav 'social sector') largely proved a failure. The effort to re-examine in principle the position of public ownership in close connection with the postulated enhancement of the role of the market (Tardos 1982), and particularly the search for institutional solutions which would effectively cut the umbilical cord linking public enterprises with state administration, may also be regarded as indicators that these are topical issues. On the other hand, by the mid-1980s there were no signs of any of the communist countries moving in the direction of pluralization of the political

system, which was regarded by some as providing a chance by which to reconcile central planning with the market mechanism (Brus 1975).

In the last quarter of the 20th century market socialism remains an active issue not only in the context of economic reform in communist countries, but also in the broader context of reappraisal of the validity of the socialist idea in general, faced with the growing challenge of new realities and new attitudes (Nove 1983).

See Also

- ▶ Command Economy
- ▶ Control and Coordination of Economic Activity
- ▶ Economic Calculation in Socialist Countries
- ▶ Lange–Lerner Mechanism
- ▶ Market Failure
- ▶ Planned Economy
- ▶ Planning
- ▶ Prices and Quantities
- ▶ Socialist Economies

Bibliography

Bajt, A. 1982. O nekim otvorenim pitanjima drustvene svojine (On some open questions of social property). *Pregled* 72(11–12): 1345–1380.

Baran, P. 1952. National economic planning, Part 3: Planning under socialism. In *A survey of contemporary economics*, vol. 2, ed. V.B. Haley. Homewood: R.D. Irwin.

Barone, E. 1908. *Il Ministerio della Produzione nella stato collectivista* (Ministry of Production in a Collectivist State). Trans. in *Collectivist economic planning,* ed. F.-A. Hayek, 245–290. London: George Routledge & Sons, 1935.

Bergson, A. 1967. Market socialism revisited. *Journal of Political Economy* 75(5): 655–673.

Brus, W. 1961. *Ogólne problemy funkcjonowania gospodarki socjalistycznej* (General problems of functioning of a socialist economy). Trans. in *The market in a socialist economy,* ed. A. Walker. London: Routledge & Kegan Paul, 1972.

Brus, W. 1975. *Socialist ownership and political systems*. London: Routledge & Kegan Paul.

Communist International. 1929. *Programme of the communist international*. London.

Dickinson, H.D. 1933. Price formation in a socialist community. *Economic Journal* 43: 237–250.

Dobb, M. 1939. A note on savings and investment in a socialist economy. *Economic Journal* 49: 713–728.

Erlich, A. 1960. *The Soviet industrialization debate*. Cambridge, MA: Harvard University Press.

Granick, D. 1975. *Enterprise guidance in Eastern Europe. A comparison of four socialist economies*. Princeton: Princeton University Press.

Hayek, F.A. (ed.). 1935. *Collectivist economic planning*. London: George Routledge & Sons.

Hayek, F.A. 1940. Socialist calculation: The competitive solution. *Economica* 7(26): 125–149.

Heimann, E. 1932. *Sozialistische Wirtschafts- und Arbeitsordnung*. Potsdam: A. Protte.

Kautsky, K. 1922. *Die proletarische Revolution und ihr Programm*. Stuttgart/Berlin: I.H.W. Dietz Nachfolger/Buchhandlung Vorwärts.

Kornai, J. 1980. *Economics of shortage*. Amsterdam/New York/Oxford: North-Holland.

Landauer, C. 1931. *Planwirtschaft und Verkehrswirtschaft* (Planned economy and exchange economy). Munich/Leipzig: Duncker & Humblot.

Lange, O. 1936–7. On the economic theory of socialism. In *On the economic theory of socialism*, ed. O. Lange, F. Taylor, and B. Lippincott. Minneapolis: University of Minnesota Press, 1948.

Lange, O. 1947. Przedmowa do polskiego wydania (Preface to the Polish edition). In *Dzieta (Collected works)*, vol. II. Warsaw: Panstwowe Wydawnictwo Ekonomiczne, 1973.

Lange, O. 1965. The computer and the market. In *Socialism, capitalism and economic growth. Essays presented to Maurice Dobb*, ed. C. Feinstein. Cambridge: Cambridge University Press, 1967.

Lerner, A. 1934. Economic theory and socialist economy. *Review of Economic Studies* 2: 51–61.

Lerner, A. 1936. A note on socialist economics. *Review of Economic Studies* 4: 72–76.

Lerner, A. 1937. Statics and dynamics in socialist economics. *Economic Journal* 47: 253–270.

Ljubljana Programme. 1958. *Program Saveza Komunista Jugoslavije* (The programme of the League of Communists of Yugoslavia). Belgrade.

Lydall, H. 1984. *Yugoslav socialism. Theory and practice*. Oxford: Clarendon Press.

Marx, K. 1867. *Capital*, vol. I. London: Progress, 1970.

Mihailovič, K. 1982. *Ekonomiska stvarnost Jugoslavije* (The economic reality of Yugoslavia). Belgrade.

Nove, A. 1983. *The economics of feasible socialism*. London: George Allen & Unwin.

Schumpeter, J.A. 1942. *Capitalism, socialism and democracy*, 5th ed. London: George Allen & Unwin, 1976.

Tardos, M. 1982. Development program for economic control and organization in Hungary. *Acta Oeconomica* 28(3–4): 295–316.

Taylor, F. 1929. The guidance of production in a socialist state. In *On the economic theory of socialism*, ed. O. Lange, F. Taylor, and B. Lippincott. Minneapolis: University of Minnesota Press, 1948.

von Mises, L. 1920. Die Wirtschaftsrechnung im sozialistischen Gemeinwesen. In *Collectivist economic planning*, ed. F.A. Hayek. London: George Routledge & Sons, 1935.
Ward, B. 1958. The firm in Illyria. *American Economic Review* 48(4): 566–589.
Wiles, P. 1969. *Communist international economics*. Oxford: Basil Blackwell.

Market Structure

John Sutton

Abstract

The term 'market structure' relates to the number and size distribution of firms in a market. Markets dominated by a few large firms are said to be 'concentrated'. This article offers a brief review of the modern literature that sets out to explain differences in concentration levels across different industries.

Keywords

Advertising; Barriers to entry; Endogenous sunk costs; First mover advantages; Herfindahl index; Integer effects; k-firm concentration ratio; Learning; Market structure; Mergers; Minimum efficient scale; Monopolistic competition; Monopoly; Multi-stage games; Network effects; Oligopoly; Perfect competition; Rate of return; Research and development; Scale economies

JEL Classifications

D4

Why is the world market for large commercial jet aircraft dominated by just two firms, while oil tankers are produced by a large number of firms spread over many countries? This is the kind of question addressed in the literature on 'market structure', a field once seen as a rather arcane area, in which explanatory theories were weak and in which discussion tended to focus on rival interpretations of 'statistical regularities' reported in empirical studies. The most famous of these 'regularities' related to a supposed link, across different industries, between the degree to which the industry was dominated by a few large firms ('concentration'), and some average measure of the rate of return (profit) on fixed assets enjoyed by firms in the industry. (Popular summary measures of concentration include the 'k-firm concentration ratio', that is, the share of industry sales revenue accounted for by the top k firms, and the Herfindahl index, defined as the sum of squares of all firms' market shares.) Now the presence of a (positive) relation of this kind would raise the question, 'why do industries with high rates of profit not attract entry, to the point where such differences are eroded?' This question was countered in the older literature, following Bain (1956), by appealing to the supposed existence of 'barriers to entry' in various industries. These barriers fell into three categories. The first related to factors intrinsic to the industry's methods of production ('scale economies'). If the average cost of production falls sharply as output rises to a certain level, then we might regard that level as a 'minimum efficient scale', and postulate that the industry is large enough to accommodate only a small number of firms of this size. This point was, and remains, uncontroversial. The second category related to institutional barriers associated with legal or regulatory impediments, or poor access to financial markets, and so on, but, while barriers of this category may be important in some industries and for some countries, they are probably of secondary relevance to the general run of industries in market economies. The third type of barrier related to the role played by advertising and R&D, and it is here that some serious difficulties arise, a point to which we turn in what follows.

The series of ideas just set out came to be known as the Bain paradigm, or the Structure–-Conduct–Performance paradigm. Expressed briefly, this view held that a more concentrated structure, however sustained, allowed firms to operate less intensive forms of price competition ('conduct'), and this in turn led to high profits ('performance'). This view was seriously undermined in the 1980s as a result of two

developments in the literature. The first of these developments was empirical: it became clear, in the light of new empirical studies, that the claim for a positive relationship between concentration and profitability was not well-founded. (For a review of the evidence, see Schmalensee 1989.) The second development was theoretical: it was clear that any successful explanation of differences in concentration across industries could not rely solely on 'scale economies' and 'institutional barriers'; the role played by advertising and R&D in raising the stakes required of entrants to an industry seemed crucial. But here a problem arises: the levels of advertising and R&D, unlike the degree of scale economies, are matters that are under the control of the firms themselves. The levels of expenditure firms undertake in these areas are ground out as part of the competitive process – and so we cannot treat their levels as a given, and claim that, when we observe a high ratio of advertising and/or R&D to industry sales revenue, this constitutes a 'barrier to entry' that explains the industry's high level of concentration. Rather, an explanation of market structure must explain *both* the level of concentration *and* the levels of advertising and R&D intensity. The 'given' that distinguishes one industry from another must not be the observed (or 'equilibrium') *level* of advertising or R&D, but rather the underlying (industry-specific) relationship between any firm's level of spending on these fixed outlays and the resulting benefit ('perceived product quality', say, or, more generally, any effect leading to an outward shift in the firm's demand schedule or a fall in its unit cost of production).

These problems with the older literature led from the late 1980s onwards to the development of a new literature on market structure. (See, for example, Dasgupta and Stiglitz 1980; Shaked and Sutton 1986; Sutton 1991, 1998. A full technical review of the literature will be found in Sutton 2007.) The point of departure of this literature lies in modelling the evolution of structure by reference to a 'free entry' model, in which any one of a number of potential entrants is free to enter the industry, and to choose its level of outlays on advertising, R&D, and so on, in the light of the choices made by its rivals.

The Modern Game-Theoretic Literature

The models used in the modern literature take the form of 'multi-stage games'. In the simplest example, a firm decides, at stage 1, to enter (and pay some positive, minimal, entry fee whose size is a given, and which can, for example, be interpreted as the cost of building a production plant of 'minimum efficient scale'). At stage 2, each firm, knowing the number of firms that have entered, chooses its level(s) of advertising and/or R&D. Its choices will depend inter alia on the (industry-specific) degree of effectiveness of these expenditures in influencing consumer demand for its product(s). In 'commodity' type industries, where this effectiveness is very low, these outlays will be close to zero. Finally, in stage 3, firms compete in price, taking as given the attributes of their respective products, and they realize corresponding levels of (gross) profits. (It is always assumed that firms have constant marginal costs of production, and that they face downward-sloping demand schedules.)

The central idea that emerges from these ('endogenous sunk costs') models is as follows: as the size of the market increases (in the sense of having a larger number of consumers, so that each firm's demand schedule shifts outwards) the industry may adjust to this in two ways: the number of firms may rise ('entry'), and/or the spending level per firm on advertising, R&D, and so on, may rise – because, in the absence of a proportional rise in the number of firms, each firm now enjoys a higher level of demand, and so the marginal return it gains from being able to charge a given price premium for a higher-quality product rises ('escalation'). Now the degree to which one or other of these effects operates depends inter alia on the effectiveness of advertising and/or R&D. It also depends on the degree to which high-quality products can draw customers away from rival products of a lower quality. Suppose, for example, that products differ not only in quality, but in other attributes also, and that customers differ in their preferences over these latter attributes. Then it will be correspondingly harder for a firm that raises its 'perceived quality' level to attract sales from rivals. An example may be helpful here:

consider, for instance, the market for flowmeters. These devices are used to measure the rate of flow of liquids, and they come in a large number of types. An increase in R&D spending by a producer of 'electromagnetic' flowmeters will have only a limited impact in drawing consumers away from 'ultrasonic' flowmeters, since the latter type of meter has attributes that makes it better suited than the electromagnetic type in certain applications. By way of contrast, consider the case of the (civil) aircraft industry, as it developed since the late 1920s. At that period, there were many types of plane in operation (monoplanes/biplanes, metal/wood construction, land/seaplanes, and so on). Yet all makers faced a market where all buyers (airlines) sought to achieve the same objective: to minimize the carrying cost per passenger mile. As soon as it became clear which type of design best achieved this single aim, plane-makers converged on the solution (an all metal monoplane with a cantilever wing design, following the Douglas DC3). Thereafter, technical developments were focused on pushing forward the performance of this type of plane, and, as plane-makers escalated their efforts in this direction, the stakes required to keep up with rival firms' innovations rose, and there was a 'shake-out' of all but a handful of firms. This story was repeated at the dawn of the jet age in the 1950s. Here a growing world market led, not to the entry of new plane-makers, but to an increasing flow of development outlays by the surviving firms, so that only Boeing and Airbus remain in the wide-body commercial jet business today. (For the details of this story, and the rise of Airbus, see Sutton 1998, ch. 15.)

Where does this leave us? The kind of market profiles that emerge are these: (*a*) those with high R&D outlays and high global concentration (for example, wide-body commercial jets); (*b*) those with high R&D outlays, low concentration and a fragmented set of distinct product categories (for example, flowmeters); and (*c*) those with low R&D spending, where, once the size of the market is large, the level of concentration may become arbitrarily low.

Within advertising-intensive industries, a simpler picture emerges. Here, the fact that a firm can use a single brand to span a range of product types in a market means that we can, with few exceptions, define markets in a way that avoids the complication posed by the presence of sub-markets for distinct product types, of the kind we encountered in the flowmeter example above. Here, the theory leads to a very simple prediction: if we take a cross-section of markets of different sizes (by looking, say, at a single industry across a number of countries of different sizes), then we will find a sharp difference in the market size–concentration relationship as between the 'advertising-intensive' industries and a control group of industries in which advertising plays an insignificant role. In the latter group, very low levels of concentration may be reached as market size increases. In the former group, concentration levels will necessarily remain above some critical level in all countries, no matter how large the size of the country (the 'non-convergence' property). This prediction, and related predictions of the theory, have been widely tested over the past decade and appears to be closely in line with what is found in the data (for a review, see Sutton 2007).

One further comment is called for in relation to this prediction: what is predicted is not an actual or equilibrium level of concentration, but rather a *lower bound* to the level of concentration that can emerge under given circumstances. It is intrinsic to models of this kind that a range of different outcomes is possible, depending on such factors as the form of the entry process (simultaneous, sequential, and so on). The most graphic illustration of this point comes from thinking about the pattern of ownership of plants spread over some geographic region large enough to support many plants. There will be a 'fragmented' equilibrium in which every plant is owned by a different firm, and there will be other equilibria in which the number of plants will remain (roughly) the same, but several of these plants will be owned by the same firm. In other words, a range of outcomes can arise as equilibria, depending on the form of the entry process and the nature of price competition, and the theoretical focus of interest lies in asking, not about the actual outcome, but about the range of possible outcomes, or, more specifically, about the lower bound to the level of concentration that can emerge.

Extensions: Learning Effects and Network Externalities

Two further ('dynamic') mechanisms play a role in explaining high levels of concentration. First, if each firm's unit cost level falls over time as a function of its cumulated volume of output to date, then an early entrant may build a dominant market position by setting an initial low price – possibly below its current unit variable cost of production – with a view to achieving a high output volume, and so a relatively low level of unit cost in the future. In a small number of industries – aircraft, semiconductors and chemical fibres – this effect is quite large.

Second, if the attractiveness of a firm's product to new consumers increases with the number of consumers it has supplied in the past, then again a firm may use an initial low price to build up its early client base and so stimulate future demand (Katz and Shapiro 1985). Examples of such effects abound in the information technology sector: as an item of hardware becomes more widely owned, more firms in the software industry will find it attractive to develop dedicated software for it, thus reinforcing its initial popularity.

What both these examples have in common with the endogenous sunk cost models described earlier becomes clear once we interpret the 'planned losses' incurred in the initial phase as a fixed outlay – analogous to an outlay on R&D or advertising – which yields a payoff in the later phase, either through lower unit costs or increased demand. The novel element which arises in these 'learning' or 'network effects' models is that these effects can be cumulative over time, and so a small initial disparity in the costs or sales of two firms may in principle become amplified over time.

Structure, Conduct and Performance Revisited

What does this imply for the Structure–Conduct–Performance paradigm? If high concentration is merely the natural outcome of the competitive process, should we still see high concentration in an industry as an indication that policy intervention might be warranted?

At a conceptual level, what remains is this: it is still true within the modern 'free entry' models discussed above that structure affects conduct. It is also true that conduct affects performance; but now there is a feedback loop through which high levels of profit may attract new entry, that is, structure is not a given, but is now determined as part of the market process. One consequence that emerges from this is that there is no simple and general link, of the kind central to the old literature, between high concentration and high profitability: it is possible, for example, to have industries with widely different levels of concentration that exhibit no difference in their rates of return on investment. High ('supernormal') profits can, however, arise in these 'free entry' models, through a number of channels. Most notably, they may arise because of asymmetries in the entry process ('first mover advantages'): an early entrant to the market may build up a level of investment in R&D, for example, and so enjoy both a high market share and a high rate of return on its investments, so that the industry-wide levels of concentration and profitability are relatively high. A second channel relates to the important but neglected role of 'integer effects', that is, if there is room in market at equilibrium for only a small number of entrants, then it may be, for example, that two firms can both make supernormal profits, but the entry of a third firm would drive the profit rate below a normal rate of return, so further entry does not occur. Finally, and most importantly, variations in productivity (unit costs) across firms associated with non-imitable advantages can lead to positive (supernormal) profits for (all) intra-marginal firms – a free entry condition implies 'zero profits' only for the marginal entrant (Demsetz 1973).

One key issue remains: what of comparisons between alternative forms of market structure within any one industry? This is the question that lies at the heart of competition policy regarding mergers. As we have seen, the normal workings of the competitive process fix a lower bound to the level of concentration that must come about under free competition. This bound can, in the

case of some industries, be very high in absolute terms, even in a large market; but in other industries it will be very low. Above this bound, varying levels of concentration can emerge since various patterns of market structure can be sustained as equilibria (as in the example of geographically dispersed plants mentioned above). What remains true of the Structure–Conduct–Performance story is that these different market structures may have different welfare properties; a proposed merger that moves us towards a more concentrated structure which will lead to reduced consumer welfare will be subject to the traditional objections. On the other hand, it may be that a merger arises merely as a response to changes in external conditions, and represents a shift away from a form of market structure that constituted an equilibrium outcome under the previous setting, but is no longer sustainable as an equilibrium in this changed environment. Distinguishing between these two possibilities in any specific instance is one of the (many) challenges in dealing with merger cases.

Why Does it Matter?

The traditional rationale for studying market structure was based on its link to profitability and to social welfare. There is, however, another line of argument that has gained considerable force as a result of the empirical success of the modern free entry models in explaining cross-industry differences in concentration. This line of argument rests on the claim that the success of these models provides convincing, though indirect, evidence for the workings of some key competitive mechanisms that appear to operate in a more or less uniform way across a wide range of industries.

To place this in perspective, it is worth noting that the conventional wisdom in economics from the 1950s to the late 1980s was deeply pessimistic in respect of models that lay between the two polar cases of perfect competition or (Chamberlinian) monopolistic competition, on the one hand, and monopoly on the other. This pessimism was typically expressed in the observation, 'with oligopoly, anything can happen'. The new game-theoretic literature of the 1980s formalized oligopoly theory using the Nash equilibrium concept, and offered it as a general framework within which perfect competition and monopoly appeared as special (limiting) cases. While this new literature appeared to some critics to simply reinforce the negative view of oligopoly theory, the successful application of these game-theoretic models to the task of 'explaining market structure' suggests that the early pessimism is unwarranted: it seems that these models capture at least some 'robust' competitive mechanisms that operate in a more or less uniform way across the general run of industries.

See Also

- ▶ Airline Industry
- ▶ Anti-trust Enforcement

Bibliography

Bain, J. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Dasgupta, P., and J.E. Stiglitz. 1980. Industrial structure and the nature of innovative activity. *Economic Journal* 90: 266–293.

Demsetz, H. 1973. Industry structure, market rivalry, and public policy. *Journal of Law and Economics* 20: 113–124.

Katz, M.L., and C. Shapiro. 1985. Network externalities, competition and compatibility. *American Economic Review* 75: 424–440.

Schmalensee, R. 1989. Inter-industry studies of structure and performance. In *Handbook of industrial organization*, ed. R. Schmalensee and R. Willig, vol. 2. Amsterdam: North-Holland.

Shaked, A., and J. Sutton. 1986. Product differentiation and market structure. *Journal of Industrial Economics* 36: 131–146.

Sutton, J. 1991. *Sunk costs and market structure*. Cambridge, MA: MIT Press.

Sutton, J. 1998. *Technology and market structure*. Cambridge, MA: MIT Press.

Sutton, J. 2007. Market structure: Theory and evidence. In *Handbook of industrial organization*, ed. M. Armstrong and R. Porter, vol. 3. Amsterdam: North-Holland.

Market Structure and Innovation

Morton I. Kamien

The study of the relationship between market structure and innovation by economists is a relatively recent phenomenon that can be traced back to the mid-1950s. Prior to that time the bulk of economic analysis took the number of products and their means of production as determined exogenously, just as consumers' tastes were taken as an exogenous given. With a few exceptions, economists appeared to be unconcerned with the economic incentives that determined the pace and direction of innovation despite the fact that this activity had begun to be institutionalized about 1876, when industrial research laboratories began to be established both in the United States and in Europe. The exceptions include Taussig (1915), Hicks (1932), Galbraith (1952), and most importantly Schumpeter (1961, 1964, 1975). It was Schumpeter who argued most persuasively that it was competition through introduction of new products and methods of production that was far more important than price competition, in the long run. For it was through innovative activity that economic development that resulted in higher per capita income took place.

The importance of technical advance as a source of growth in per capital income was dramatized by Solow's (1957) claim that ninety per cent of the doubling of per capita output in the US non-farm sector in the period 1909–49 was the result of technical advance and only the remaining ten per cent was the result of an increase in the amount of capital used by each worker. Despite subsequent refinements of this estimate to take into account increases in the quality of labour and capital, technical advance remained a major source of growth in per capita output.

Recognition that technical advance was a major source of economic growth caused attention to be turned to the study of the conditions in a market economy that facilitated it and those that retarded it. These conditions had largely been outlined by Schumpeter. He contended that possession of some degree of monopoly power by a firm was necessary in order for it to engage in innovative activity. There are two reasons for this contention. The first is that the monopoly profits associated with monopoly power provided an internal source of funding of research and development. Now internal funding of research and development, as opposed to financing through borrowing, is commonplace and essential. This is because a research and development project provides little in the form of tangible collateral for lenders to recoup should it fail. Moreover, since lenders are typically not in a good position to monitor carefully the progress of a research and development effort, or would find it prohibitively costly to do so, they are confronted with a moral hazard problem in the form of potential shirking by the managers of the project. A substantial investment in the project by its managers helps alleviate this concern of its external financiers. Finally, external financing of a research and development project may require disclosure of a level of information that is incompatible with maintaining its secrecy from potential imitators.

The second reason that some monopoly power is important for innovative activity is that its possession enables an innovator to reap profits from his investment and thereby provides him the incentive for undertaking it in the first place. The monopoly power here is in the form of the ability to prevent quick imitation and erosion of the profits from the innovation. This form of monopoly power is often embodied in a patent, a copyright, or a trademark but may also be embodied in a less formal manner such as possession of channels of distribution. To Schumpeter's contention that a degree of monopoly power was necessary for innovative activity Galbraith added the claim that large firm size was also essential. It was his claim that only large firms had adequate resources to finance internally current research and development projects, as all the cheap innovations, that could be undertaken by individuals, had already been done. Moreover, large firms could diminish the overall risk of research and development

activity by engaging in a large number of different, uncorrelated projects.

The hypotheses that have emerged from Schumpeter and others can be summarized as follows:

(1) Innovation is greater in monopolistic industries than in competitive ones because
   (a) a firm with monopoly power can prevent limitation and thereby can capture more profit from an innovation;
   (b) a firm with monopoly profits is better able to finance research and development.
(2) Large firms are more innovative than small firms because
   (a) a large firm can finance a large research and development staff. There are economies of scale in this activity also;
   (b) a large diversified firm is better able to exploit unforeseen innovations;
   (c) indivisibility in cost-reducing innovations makes them more profitable for large firms.
(3) Innovation is spurred by technological opportunity.
(4) Innovation is spurred by market opportunity.

Almost all the facets of these hypotheses have been subjected to some degree of empirical investigation. The leading figures in this effort include Griliches (1957, 1984), Mansfield (1968a, b, 1971, 1977) and Scherer (1980), as well as many others. Extensive surveys of this work can be found in Kamien and Schwartz (1982) and Stoneman (1983). Roughly speaking these investigations reveal that innovative activity is most intense in industries with a market structure intermediate between a perfectly competitive market and a perfectly monopolistic one. Too much competition appears to discourage innovative activity as innovators are unable to capture enough of the rewards from it. On the other hand, too much monopoly power appears to lead to complacency and less innovative activity. Large firms do spend more on research and development and obtain more patents than small firms in absolute terms but not proportionately more as measured by their market share, total sales, or value of assets except in the chemicals industry, especially pharmaceuticals manufacturers. There appears to be a threshold level of firm size below which firms do not engage in formal research and development efforts and above which they do. Research and development activity appears to grow proportionately with firm size up to a point and then grow less than proportionately as firm size increases beyond this point. Research and development activity appears to be more efficient in medium size firms than in large firms, as measured by the cost of developing a patent. Technological opportunity does spur innovative activity in an industry but so does market opportunity, in the form of demand for new products and methods of production. Research and development projects that are undertaken in response to market demand appear to succeed more than those that are undertaken as a result of technological opportunity.

These empirical findings have spawned a new generation of theoretical models that can be traced to Scherer (1967), Brazel (1968), Kamien and Schwartz (1972, 1976), Loury (1979), Dasgupta and Stiglitz (1980a, b), and Reinganum (1981, 1982). A survey of these models can be found in Kamien and Schwartz (1982) and a more recent overview in Reinganum (1984). These models have come to be referred to as patent race models and they constitute an area of intense research activity within the field of industrial organization. The continued interest in the economies of technical advance stems, of course, from the fact that technical advance is perhaps the most important determinant of our past, our present, and our future.

See Also

► Entry and Market Structure
► Innovation

Bibliography

Brazel, Y. 1968. Optimal timing of innovations. *Review of Economics and Statistics* 50: 348–55.

Dasgupta, P., and J. Stiglitz. 1980a. Industrial structure and the nature of innovative activity. *Economic Journal* 90: 266–93.

Dasgupta, P., and J. Stiglitz. 1980b. Industrial structure and the speed of R&D. *Bell Journal of Economics* 11: 1–28.
Galbraith, J.K. 1952. *American capitalism*. Boston: Houghton Mifflin.
Griliches, Z. 1957. Hybrid corn: An exploration of the economics of technological change. *Econometrica* 25: 501–22.
Griliches, Z. (ed.). 1984. *R&D, patents, and productivity.* Chicago: University of Chicago Press.
Hicks, J.R. 1932. *The theory of wages*. London: Macmillan.
Kamien, M.I., and N.L. Schwartz. 1972. Timing of innovations under rivalry. *Econometrica* 40: 43–60.
Kamien, M.I., and N.L. Schwartz. 1976. On the degree of rivalry for maximum innovative activity. *Quarterly Journal of Economics* 90: 245–60.
Kamien, M.I., and N.L. Schwartz. 1982. *Market structure and innovation*. Cambridge: Cambridge University Press.
Loury, G.C. 1979. Market structure and innovation. *Quarterly Journal of Economics* 93: 395–410.
Mansfield, E. 1968a. *The economic of technological change*. New York: Norton.
Mansfield, E. 1968b. *Industrial research and technological innovation; an econometric analysis*. New York: Norton.
Mansfield, E., et al. 1971. *Research and innovation in the modern corporation*. New York: Norton.
Mansfield, E., et al. 1977. Social and private rates of return from industrial innovations. *Quarterly Journal of Economics* 91(2): 221–40.
Reinganum, J.F. 1981. Dynamic games of innovation. *Journal of Economic Theory* 225: 21–41.
Reinganum, J.F. 1982. A dynamic game of R and D: Patent protection and competitive behavior. *Econometrica* 50: 671–88.
Reinganum, J.F. 1984. Practical implications of game theoretic models of R&D. *American Economic Review* 73: 61–66.
Scherer, F.M. 1967. Research and development resource allocation under rivalry. *Quarterly Journal of Economics* 81: 359–94.
Scherer, F.M. 1980. *Industrial market structure and economic performance*. Chicago: Rand McNally.
Schumpeter, J.A. 1961. *Theory of economic development*. New York: Oxford University Press.
Schumpeter, J.A. 1964. *Business cycles*. New York: McGraw-Hill.
Schumpeter, J.A. 1975. *Capitalism, socialism and democracy.* New York: Harper & Row, Colophon Edition.
Solow, R.M. 1957. Technical change and the aggregate production function. *Review of Economics and Statistics* 39(August): 312–20.
Stoneman, P. 1983. *The economic analysis of technological change*. Oxford: Oxford University Press.
Taussig, F.W. 1915. *Investors and money-makers: Lectures on some relations between economics and psychology.* New York: Macmillan.

Market Value and Market Price

Anwar Shaikh

Marx defines the labour value of a commodity as the total (direct and indirect) abstract labour time required for its production. It is his contention that under capitalism the movements of commodity prices are dominated by changes in labour value magnitudes. This thesis, which he calls the law of value, requires him to connect labour values to the different *regulating prices* which act as centres of gravity of market prices under various assumed conditions of production and sale. He therefore undertakes to systematically develop the category of regulating price by introducing successively more complex factors into the analysis, linking it at each step to its foundation in labour value. It is only near the end of this developmental chain, when he begins to analyse the manner in which differences among conditions of production within an industry influence the process of regulating market prices, that we encounter the concept of *market value* (Marx 1894, ch. X). To grasp its significance, we must first consider the steps which precede it.

The simplest expression of the law of value is when exchange is directly regulated by labour values. If we define direct price as a money price proportional to a commodity's labour value, then the simple case corresponds to the situation in which the direct price of a commodity is the regulating price (i.e. centre of gravity) of its market price. Marx begins with this premise in Volume I of *Capital*, concretizes it in Volume II to account for turnover time and circulation costs, transforms it in Volume III into the notion of prices of production (prices reflecting roughly equal rates of profit) as regulating prices, and then goes on to develop even this concept further, to account for rental payments, trading margins and interest flows. It is important to note that throughout this whole process of developing the various forms of regulating price, the aim is not only to encompass the complexity of the

determinants of market prices, but also to show their connection to labour values.

The above path focuses on the complex character of the centres of gravity of various types of market prices. But the very concept of a gravitational centre itself requires some discussion of the forces of supply and demand, because it is through their variation that the market price of a commodity is made to orbit around its (generally moving) centre of gravity. Accordingly, Marx also engages in a second, parallel, discussion of the manner in which a regulating price exerts its influence over market price. And here, the basic idea is that when (for instance) the growth of demand exceeds that of supply, market price will rise above regulating price, and the resulting rise in profitability above its regulating level (as embodied in the assumed regulating price) will induce capitalists to accelerate supply relative to demand. The original gap between supply and demand will thereby be reduced or even reversed, thus driving the market price back towards or even below the regulating price. In this way, the dynamic adjustment of supply to demand serves to keep market price oscillating around the regulating price. Note that the whole argument is cast in terms of the relative *growth rates* of supply and demand rather than merely in terms of their (implicitly static) levels, and that market prices continually oscillate around regulating prices without ever having to converge to them in any mythical 'long run equilibrium' (Shaikh 1982).

The preceding analysis implicitly ignores any variations in unit production costs and unit labour values, so that the regulating price itself is assumed to be unchanged during the regulation process. This is adequate as long as we abstract from differences among conditions of production within a given industry, because then each individual producer in effect embodies the average conditions and the whole story can be told simply in terms of the average producer. Under these circumstances, it is the *social* (i.e. average) unit labour value which ultimately regulates the movements of market prices, through the mediation of a particular regulating price. As Marx puts it, it is the social value of the commodity which functions here as the labour value which is regulative of market price, i.e. as the *market value*.

The obvious next step is to introduce the issue of differences among producers within an industry. Accordingly, Marx examines the situation in which there are three types of production conditions in use, ranked in order from lowest efficiency (1), to medium (2), to best (3). The ranking of individual unit labour values (and unit production costs, other things being equal) will of course be in reverse order. As before, the social unit value is the total labour value of the total product divided by the amount of this total product. But this average now represents not only 'the average [unit] value of commodities produced' in this industry, but also the unit 'individual value of the ... average conditions' in the industry. Note that although the unit social value will be 'midway between the two extremes', it can nonetheless differ from the medium (2) unit value precisely because the average of existing conditions can differ from the medium (2) condition according to the weights of low (1) and high (3) conditions in total output.

The important thing at this juncture is to identify the specific conditions of production which operate to regulate market price through the ebb and flow of supply, *because it is the labour value of these particular conditions which will therefore function as the market value*. This leads him to identify three types of response to a deviation of market price from some pre-existing regulating price. The first case is when all three conditions of production are able to adjust their respective rates of supply, so that the average production condition continues to regulate the market. Here, the regulating price still rests upon the average unit production cost, and the unit social value is still the market value. The only new consideration is that the regulating price and market value may vary within certain strict limits, because the functioning average condition of production may itself change insofar as the weights of its three constituent types of production conditions alter over the adjustment process. To the extent that better conditions accelerate more in the up phase and worse conditions decelerate more on the down side, even this effect will more or less cancel out

over a given oscillation of market price around regulating price.

At the other extreme, Marx considers situations where the deviation of market price from regulating price goes so far as to bring either the worst or best production to the fore as the foundation of new regulating prices and market values. It is plausible, for instance, that the utilization of capacity is usually inversely correlated with the efficiency of production. Then, if demand rises sufficiently, the bulk of the slack will be taken up at first by the best, then by the intermediate, and finally by the worst conditions of production. A situation may therefore arise in which the unit production costs of the worst conditions of production will come to determine the regulating price, so that the individual unit labour value of these conditions becomes the market value. Conversely, a sufficiently rapid fall in demand relative to supply may precipitate just the opposite situation, in which only the best conditions survive to regulate the market price and thus determine the regulating price and market value. It should be noted, incidentally, that while the shift of regulating conditions to one extreme or the other is precipitated here by 'extraordinary combinations' of supply and demand, this need not be the case when we consider technical change (in which the regulating conditions will be the best *generally accessible* methods of production) or production in agriculture and mining (in which the regulating conditions are often the ones on the margin of cultivation and location, hence among the worst of the lands and locations in use). From this point of view, Marx's initial discussion of Market Value is merely prelude to the much broader question of regulating value and conditions of production.

See Also

- ▶ Market Price
- ▶ Marxian Value Analysis

Bibliography

Itoh, M. 1980. *Value and crisis: Essays on marxian economics in Japan*. London: Pluto Press, ch. 3.

Marx, K. 1894. *Capital*, Vol. III, ch. X. New York: International Publishers, 1967.

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

Marketing Boards

Christopher B. Barrett and Emelly Mutambatsere

Abstract

Marketing boards (state-controlled or state-sanctioned entities legally granted control over the purchase or sale of agricultural commodities) flourished in the 20th century. Since the mid-1980s they have declined in number under pressure from domestic liberalization and from international trade rules that increasingly cover agriculture. Where reforms have been widespread and successful, marketing boards have vanished or retreated to providing public goods, such as strategic grain reserves or insurance against extraordinary price fluctuations. Elsewhere, the weaknesses of private agricultural marketing channels have been revealed by the rollback of marketing boards, often leading to calls for reinstatement of powerful marketing boards.

Keywords

Agricultural markets in developing countries; Agriculture in economic development; Contract enforcement; Corruption; Deregulation; Great depression; International monetary fund; Lomé Convention; Marketing boards; Monopoly power; Monopsony power; Price control; Price discrimination; Price stabilization; Privatization; Protection; Structural adjustment programmes; Subsidies; World Bank; World Trade Organization

JEL Classifications

Q1

Marketing boards are state-controlled or state-sanctioned entities legally granted control over the purchase or sale of agricultural commodities. They can be divided into two broad categories. Monopolistic marketing boards that create a single-commodity seller are found mainly in developed countries. Monopsonistic marketing boards concentrating buyer-side market power in one institution were commonplace for many years in developing countries. Monopolistic marketing boards were typically established with the main objective of maintaining or raising and stabilizing farm prices and incomes in an administratively practical and politically acceptable manner. By contrast, monopsonistic marketing boards were typically established to give the state control over commodity prices – normally for the benefit of foreign and urban buyers – and capacity to tax agriculture so as to subsidize industrialization.

Marketing Boards in Developed Countries

Marketing boards are state-sponsored trading enterprises legally invested with monopoly powers to organize the marketing of agricultural commodities. These statutory entities typically operate under direct or indirect producer control. Among the earliest boards were the New Zealand Meat Producers Board and the New Zealand Dairy Board, each established in 1922, the Australia Queensland Sugar Board of 1923, and the Australia Wheat Board, formed in 1939. In Australia, marketing boards used import protection and home consumption price schemes to stabilize producer prices. They initially received financial support from the state, although such support later declined as the focus of the boards changed. A number of state and commonwealth-level marketing boards were later established, with varying degrees of authority and responsibilities in the marketing of agricultural products such as wool, dairy, meat, wine and brandy, honey and horticultural products. The marketing boards in New Zealand evolved in a similar manner, with regulatory authority in export marketing and licensing but no direct financial support from the state. These boards, involved in the marketing of dairy, apple and pear, kiwi fruit, horticulture, meat and wool products, all used activities such as single-desk selling, price pooling, revenue pooling and preferential financing to seek higher producer prices.

The earliest major marketing schemes in Britain were the milk, potatoes and bacon marketing boards formed under the British Marketing Acts of 1931 and 1933. These acts enabled producers to set up marketing schemes that had the legislative power to ensure conformity by all producers. The core purpose of the marketing boards was to maintain or raise producer prices of basic agricultural commodities through acreage restrictions, direct or indirect limits on saleable quantities, and price discrimination, with higher prices in sheltered markets and lower prices in exposed ones. In addition, monopolies of processed products were legalized, leading to the organization of processor and distributor schemes. The marketing boards thus held the monopoly power to control supply, the terms of sale and the channels and conditions of sale (Bauer 1948). By 1948 marketing boards had spread to include all major agricultural commodities. In Canada, marketing boards were also formed in response to the price fluctuations of the Great Depression. The Dominion Marketing Board, a federal agency established under the National Farm Products Act of 1934, exercised extensive market power over the sale of regulated products, transferable to provincial-level producer-organized boards. The Agricultural Marketing Acts of 1940 and 1956 delineated the powers of regulation and market control activities for the established and new federal and provincial marketing boards. The result was marketing boards with diverse market powers and scope of operations across provinces, and across boards within the same province. Some marketing boards act only in a supervisory capacity, whereas others wield more extensive powers in market regulation and control. Activities generally range from negotiating minimum prices, regulating quantity and quality of marketed products, collecting and distributing payments, as well as grading and quality control.

Several common features distinguish marketing boards in developed countries from those

found in developing nations. First, marketing boards in developed countries tend to be specialized in both scale and scope of operations. For example, New Zealand currently runs strictly export monopolies, such as the Dairy Board, that have control over the country's agricultural exports but negligible influence over domestic production, sales, imports or tariff rates.

Second, marketing boards in developed countries tend to subsidize farmers at the expense of consumers, as evidenced by their mandate to maintain high producer prices for farmers through limited supply. One result is that marketing boards in developed countries have tended to generate windfall profits for the owners of farm land and other sector-specific assets in agriculture.

Third, and following directly from their role in subsidizing farmers, state trading enterprises tend to encourage and support cartels at producer, processor and distributor levels. Developed country agricultural marketing boards have been a major issue in international trade because historically they dominated certain markets. For example, McCalla and Schmitz (1982) estimated that 95 per cent of world wheat trade in 1973–7 involved a state marketing board on at least one side of the transaction. Because marketing boards enjoy greater flexibility than private traders in pricing – for example, they can commonly delay payments to producers, pool payments so as to reduce producer price risk, and can practise discriminatory pricing among export or import markets – their operations are closely scrutinized by the World Trade Organization for prospectively anti-competitive practices.

Marketing Boards in Developing Countries

Marketing boards in developing countries were typically begun during colonial times for purposes distinct from those of their counterpart marketing boards in developed economies. And they have followed a somewhat different trajectory from those of marketing boards in developed countries.

European colonial powers formed marketing boards in large measure to facilitate the export of agricultural commodities to Europe and to stabilize prices faced by colonial elites (for food crops) and metropolitan buyers (for export crops). Post-independence governments generally maintained marketing boards because these were considered simpler to manage and more efficient in conducting organized trade than the traditional, decentralized private sector. More compellingly, marketing boards provided a convenient way for the governments to maintain control over the marketing of strategic commodities, such as the food staples and important export crops (Lele and Christiansen 1989). The marketing boards system was most prevalent in the anglophone African and South Asian countries, but widespread as well in francophone and lusophone African countries and in Asia and Latin America.

Marketing boards were both state-owned and state-funded, based on centralized decision making systems. They possessed the sole legal authority to purchase commodities from farmers and to engage in trade. Through the boards, governments typically fixed official producer prices for all controlled commodities, often in a pan-seasonal and pan-territorial manner whereby a single price was set for the whole marketing season and for all regions of the country. Marketing boards provided a guaranteed market for the farmers, absorbing all marketed surplus at the official producer prices, and maintaining extensive buying networks and storage facilities throughout the production regions. Pan-seasonal and pan-territorial pricing practices eliminated any opportunities for arbitrage, discouraging private investment in commodity storage or transport capacity, and reinforcing the government's control over the marketing channel. Unlike marketing boards in developed countries, producer sales into the network were rarely rationed, because the marketing boards'objective was normally to increase supply and lower prices for consumers, as opposed to controlling supply for the benefit of producers.

Two features of the export crop marketing boards – as distinct from those handling staple food commodities – are worth noting. First, the marketing boards held the sole legal rights in commodity export, and had a mandate to generate income for the state. Therefore, storage costs were

maintained at low levels through selling policies such as rapid evacuation and forward selling. In addition, local producer prices were typically set at levels lower than the international free-on-board prices, through price fixing or overvalued exchange rates. Essentially, export crop marketing boards were used as a means to tax agriculture in order to develop the industrial sector in these agrarian economies. The taxes were often quite severe. In Tanzania, for example, local producer prices for coffee and tobacco fell to 23 per cent and 15 per cent of international prices, respectively, by the mid-1980s.

Second, because export crop marketing boards served foreign demand, no price controls existed on the selling end. Marketing boards could trade on an open market for the highest possible selling prices. However, because most of the former European colonies enjoyed preferential access to European markets under the Lomé Convention, most commodities were sold to Europe. In addition, some export crops enjoyed commodity price stabilization through international commodity agreements such as the International Coffee Agreement or the International Rubber Agreement. In those cases where a country enjoys world market power, a state marketing board can, at least in theory, increase prices and thereby extract consumer surplus from foreign buyers to benefit the exporting country, including its producers. This is one of the concerns surrounding state trading enterprises within global trade policy fora.

Even though the export crop marketing boards were generally established first, in most developing countries staple food commodity marketing boards became at least as significant a part of the parastatal system. For food commodities, government control extended to every stage of the market chain, to include farm gate, wholesale and retail price controls. In-country commodity movement was restricted, especially the movement of strategic food commodities, and private trade was either illegal or legal only by licence. To achieve food security objectives, food subsidies were generally offered, mostly implicitly, in the form of fixed consumer prices set at levels lower than the market price. Although farm prices were generally set at a below-market level as well, the government often offered implicit subsidies to farmers, through price stabilization operations, and input and credit subsidies administered through the marketing boards (Lele and Christiansen 1989). Moreover, pan-territorial pricing typically implied subsidies for farmers in more remote smallholder regions. In some countries and for some crops, these arrangements likely stimulated greater crop production than would have occurred under open market arrangements.

Grain marketing boards commonly also handled the strategic food reserves for emergency situations, and had the responsibility to import food in shortage seasons. These parastatals therefore held most of their nations' inter-seasonal and inter-annual grain storage capacity, a legacy that would affect inter-seasonal commodity price movements after the liberalization of commodity marketing systems in the 1980s and 1990s. Although processing was not their core business, marketing boards, in some cases, were also involved in preliminary processing, such as milling rice or maize, or in licensing and monitoring the processing industry activities. This underscores an important difference from developed country marketing boards: the breadth of commodity marketing boards' mandate in most developing countries.

Over time, the fiscal sustainability of marketing boards in developing countries became questionable. The broad range of marketing operations handled by marketing boards and the politically charged manner in which these operations were typically handled led to massive inefficiencies and deficits that cash-strapped central governments had an increasingly difficult time covering. The subsidies embedded in grains pricing systems, coupled with heavy overhead costs associated with high administrative, transportation and storage costs, soon created huge tax burdens. In an attempt to ensure food security, the state would generally increase producer prices with less than proportional increases in consumer prices, taking on responsibility for a significant share of the marketing costs associated with moving food from farm to table. The pan-territorial pricing system meant higher transportation and handling costs in moving commodities from some remote areas, and the management of large volumes of

commodities in storage was costly. In addition, the monitoring of private trade was not only costly but generally ineffective, especially for food commodities in shortage seasons, when parallel markets flourished to meet local demand. In Mali, for example, even though private cereals trade was illegal before 1981, only 30–40 per cent of total grain trade was actually handled by the state trading agency, OPAM (Dembélé and Staatz 2002). On the international market, marketing boards faced decreasing real commodity prices for export crops, further undermining their sustainability.

By the end of the 1970s budget deficits resulting from the management and mismanagement of parastatals had reached astronomical levels in most countries. In Mali, OPAM's annual deficit reached US$80 million by 1980, three times the board's annual grain sales. In Tanzania, the National Marketing Corporation's overdrafts were about $250 million in 1993, against total state expenditures on agriculture of $12 million. The National Cereals and Produce Board (NCPB) of Kenya accumulated an estimated loss of about $300 million by 1993, in contrast with central government expenditure on agriculture of $33 million (Staaz et al. 2002; Lele and Christiansen 1989). These patterns were by no means exclusive to Africa. Indonesia's price stabilization scheme for rice, managed by the National Logistics Supply Organization (BULOG), also proved a high price to pay for self-sufficiency, as did the Food Corporation of India.

In addition to budgetary complications, marketing boards also faced organizational challenges. Their susceptibility to bureaucracy and corruption increased both the inefficiency in their operations and the transactions costs for farmers and consumers. For example, Arhin et al. (1985) argue that by the mid-1970s the Ghana Cocoa Marketing Board had become little more than an instrument of the government for the purpose of mobilizing political support for the incumbent government.

Mounting deficits, poor management and the perverse incentives created by anti-competitive behaviour brought marketing boards and price stabilization systems under attack, based in part on seminal research into the welfare effects of government interventions to stabilize commodity prices (Newbery and Stiglitz 1981). These deficit problems, coupled with the new economic insights, triggered widespread agricultural market reforms in the 1980s and 1990s throughout the developing world, implemented mainly but not exclusively, in the context of structural adjustment programmes (SAPs) of the World Bank and the International Monetary Fund.

Agricultural marketing reforms generally aimed to reduce the role of the public sector in marketing and to encourage private sector participation so as to let markets allocate scarce goods more efficiently. Marketing boards experienced major reforms under these programmes, comprising the elimination of price controls, termination of farm input and consumer food subsidies, removal of marketing boards' monopsony power and deregulation of private trade. In many cases, marketing boards were privatized or at least commercialized, the latter referring to cases where marketing boards remained government owned, but with autonomous decision-making power and an explicit objective to maximize profits. The logic was that, by removing political interference in the marketing process, market forces would lead to efficient resource allocation and price discovery. Market deregulation was thus expected to improve marketing efficiency by reducing transactions costs, increasing producer prices, thus inducing increased production and potentially also lowering consumer prices.

The response of the market was immediate and quite dramatic in many cases. Entry into formerly controlled agricultural markets was massive in most countries, although with continued bottlenecks in functions requiring significant capital outlays, such as bulk inter-seasonal storage and long-haul motorized transport, entry was typically restricted to niches with low entry barriers (Barrett 1997). Nonetheless, formal and informal private traders became a significant part of the marketing channel, performing most of the trade activities that the marketing boards previously performed.

In spite of widespread liberalization, marketing operations for most 'strategic' food and export crops changed little. Newly privatized or commercialized marketing boards were often replaced with 'new'

marketing boards that were initially intended to provide public goods, but eventually and predictably became involved in crop marketing. In Zambia, for example, the government-owned Food Reserve Agency (FRA) that replaced the National Agricultural Marketing Board (NAMBOARD) in 1995, charged with maintaining the strategic grain reserve and acting as a buyer of last resort for smallholder farmers, in time took up prior NAMBOARD responsibilities such as fertilizer distribution. Moreover, some of the commercialized marketing boards did not significantly change their pricing systems and continued to use the power of the state to remain dominant players in the current market system. In Indonesia for example, even though the market was opened to private traders, BULOG remained a price leader by operating a major buffer stock, purchasing rice when rice prices fell below a stated floor price and releasing stocks when prices rose above a price ceiling. Similarly, in the Kenyan maize sector the NCPB continued to intervene directly in markets to support maize prices; and in Malawi ADMARC remains the dominant maize buyer and distributor of inputs. Zimbabwe went so far as to reinstate the monopsony power of the Grain Marketing Board and its pre-reform operations. Not surprisingly, the budget deficit of these marketing boards actually increased after reforms.

These trends reflect governments' reluctance to relinquish control over marketing board operations, particularly the setting of prices for key food and export crops, given political sensitivity to these issues. As it turned out, such concerns were not completely unwarranted. In many developing countries the legacy of private underinvestment in storage and transport capacity, inadequate commercial trading skills in the nascent private sector, combined with limited access to finance, restricted entry into key niches of the marketing channel. These market conditions facilitated the emergence of new monopolies, often substituting private for public market power. Problems of weak contract enforcement, unreliable physical security and underdeveloped communications and transport infrastructure often impeded business expansion, market integration and price transmission. Despite increased private investment in transportation and storage infrastructure after reforms, the weaknesses of the existing systems implied considerable business risk. Consequently, private traders did not fully or quickly fill the voids left by the withdrawal of the marketing boards from core commodity market intermediation functions. Price volatility increased sharply in many countries. Moreover, farmers' access to seasonal credit dropped significantly as market liberalization ended formerly monopsonistic marketing boards' willingness to extend seasonal credit to growers that were collateralized by future sales. Reduced credit often led to fewer purchased inputs and lower crop output. In an attempt to restore market stability and production volumes, states often suspended or reversed reforms, reinstating price controls and trade restrictions, thereby further exacerbating instability and undermining investor confidence. The result has been incomplete reforms in most developing countries, where private sector involvement remains pervasive but small-scale and weak, while unprofitable commercialized marketing boards remain prominent and prone to government interference.

The Current State of Play

Far fewer marketing boards exist than previously. Because they reduce or eliminate competition, marketing boards are widely believed to induce inefficiency in marketing and sluggishness in price discovery. Therefore, government involvement in agricultural marketing has been weakening in both developed and developing countries since the mid-1980s, a result of the adoption of more liberal domestic economic policies, coupled with global pressure to conform to international trade rules steadily expanding their coverage of agriculture. The monopoly or monopsony powers of all but a few marketing boards have been lifted, and the marketing and processing activities of the boards have been streamlined. Where reforms have been widespread and successful, marketing boards have vanished or retreated to providing public goods, such as strategic grain reserves or insurance against extraordinary price fluctuations. Where reforms have been halting or unsuccessful, the weaknesses of private agricultural marketing

channels have been laid bare by the rollback of marketing boards.

See Also

- ▶ Agricultural Markets in Developing Countries
- ▶ Agriculture and Economic Development
- ▶ International Trade Theory

Bibliography

Arhin, K., P. Hesp, and L. van der Laan. 1985. *Marketing boards in tropical Africa*. Nairobi: KIP Limited.

Barrett, C. 1997. Food marketing liberalization and trader entry: Evidence from Madagascar. *World Development* 25: 763–777.

Bauer, P. 1948. A review of the agricultural marketing schemes. *Economica* 15: 132–150.

Dembélé, N., and J. Staatz. 2002. The impact of market reform on agricultural transformation in Mali. In *Perspectives on agricultural transformation: A view from Africa*, ed. T. Jayne, I. Minde, and G. Argwings-Kodhek. Hauppauge: Nova Science Publishers Inc.

Lele, U., and R. Christiansen. 1989. *Markets, marketing boards and cooperatives in Africa, issues in adjustment policy.* MADIA Discussion Paper 11. Washington, DC: World Bank.

McCalla, A., and A. Schmitz. 1982. State trading in grain. In *State trading in international markets*, ed. M. Kostecki. New York: St Martin's Press.

Newbery, D., and J. Stiglitz. 1981. *The theory of commodity price stabilization*. Oxford: Clarendon Press.

Marketplaces

Winifred B. Rothenberg

Abstract

'Marketplace' was defined by the great 18th century jurist, William Blackstone, as 'a spot of ground set apart by custom for the sale of particular goods.' The definition is striking in its simplicity, but the simplicity is deceptive, for each word should give us pause. If a marketplace is a 'spot' it is both bounded and of little consequence. Having said it was a 'spot', 'of ground' would seem to be redundant were it not that this, the market as a *place* located in space is the feature that over time will change the most. 'Set apart' underscores the irreconcilability of Commerce and Community, and, by implication, bestows upon Community the greater authenticity. 'Sale'–legally to divest a seller – presupposes a body of legal rules protecting title and the alienability of title, at least to 'particular goods'. And 'particular goods', in turn, implies the inalienability of title to *other* goods. Finally, there is 'custom'. In Blackstone's syntax, custom's role appears limited to the setting of the spot. But when Commerce threatens to overwhelm the barriers that Community has erected against it, it will be custom that writes the regulations into law. 'Marketplace' is not at all 'simple,' and reckoning with it in one or another of the protean forms it has assumed over the millennia deserves to engage, as indeed it has, the attention of archaeologists, historians, anthropologists, sociologists, and economists.

Keywords

Black Death; Blackstone, W.; Braudel, F.; Domar, E.; Durkheim, E.; Law of one price; Market institutions; Marketplaces; Marshall, A.; Non-market economies; Peasant economy; Peasants; Regulation of markets; Repeated games; Silk Road; Social networks

JEL Classifications

R11

Marketplaces in the Long-Run

The history of marketplaces is traced here as a sequence of pivotal events that began at least 60,000 years ago when, archaeologists suggest, the earliest appearance of trade in Europe can be inferred from the discovery of made objects – amber, mollusc shells and worked stone – that had been moved to sites hundreds of kilometers from their sources. But the origins of trade may extend back even further in time, for 'the

archaeological record does not reveal *any* time in the history of man in Europe in which there was no movement of "manufactured" objects' (Grantham 1997, p. 18).

A giant leap takes us to Mesopotamia 10,000 years ago. Having by now collected abundant evidence of the deleterious effect of settled agriculture on human health in the Neolithic, it has been suggested by some paleopathologists that the Agricultural Revolution – the sedentary cultivation of grain, sugar, flax, wool, sisal, jute and hemp – may have been driven less by the quest for food security than by the high value these staple crops could earn in trade.

Study of the 4,000 year-old Heqanakht Papyri reveals Heqanakht himself to have been 'obsessed' with running his farm to make a profit and augment his wealth, and the economy of Pharaonic Egypt to have been one *not* of priestly redistribution, as had been thought, but of private property, money prices, cash crops, rental land, wage labour and marketplaces (Allen 2002).

Ten centuries before the Christian era, the marketplaces rimming the Aegean Basin were specializing in the export of high-value wines, olives and oil to trade for imports of grains, raw materials and slaves from the 'barbarians' north of the Mediterranean.

In the eighth and seventh centuries BC, Phoenicia's legendary traders integrated the markets of the eastern and western Mediterranean; the Etruscans made contact with the Celts; and the merchants of India reached the northwest coast of England.

In the Bible, The Book of Ezekiel, written, it is thought, in the sixth century BC, describes in chapters 26–27 the prosperous city of Tyre whose marketplaces overflowed with an abundance of fir trees from Senir, cedars from Lebanon, ivory benches from the isle of Chittim, fine linen from Egypt, silver, iron, tin and lead from Tarshish, emeralds, purple, coral, agate and fine linen from Syria, wheat from Minnith, honey, oil and balm from Pannag, wine and white wool from Helbon, lambs, rams and goats from Arabia, spices from Sheba and Ra-amah, and skilled artisans, laborers seafaring men, and merchants from all over the eastern Mediterranean.

Most exotic of all were the marketplaces that grew up along the Silk Road, established by the Han Dynasty in the second century BC as China's only link to the West. Along much of its length, the Silk Road was less a 'road' than a hazardous path around the world's most unlivable desert, through the world's most inaccessible mountain passes, over the world's highest peaks, and among the world's most isolated and hostile peoples. The Road started in what is now Xian on China's northwest border, and made its way west through Kazakstan, Kyrgyzstan, Uzbekistan, Tajikistan, and Afghanistan, encircling the Hindu Kush, until crossing the Black Sea, it ended in Roman Syria – which is as much as to say, in Venice! The market-places of this commerce were the oases of central Asia and the fabled 'Arabian Nights' cities they became: Xian, Islamabad, Tehran, Kashi, Samarkand, Bukhara, Kabul, Kandahar, Tashkent, Aleppo, Lahore, Baghdad, Ankara, Istanbul. For 15 centuries, caravans of up to 1,000 camels each made the journey from oasis to oasis for months on end, under armed guard, bearing gold, metals, ivory, precious stones, myrrh, frankincense, ostrich eggs, horses, glass, silks, porcelain, guns, powder, jade, bronzes, lacquer – and the great religious civilizations of Buddhism and Islam.

On the European end of the Silk Road great 'diaspora networks' were founded by Greek, Armenian and Jewish merchant families of the eastern Mediterranean, who traded a world away with the diaspora networks of India, China, Japan and Indonesia (McCabe, Harlaftis and Minoglou 2005).

In time, the Romans would discover the secret of making silk, Europeans would find a sea route to China, the Silk Road with its storied past would disappear beneath the sand, the Mongol hordes would burst out of the East, and Europe would reap the whirlwind. In the Black Death that ravaged the continent between 1348 and 1351, Europe lost as much as 40 per cent of its population.

To a demographic event so catastrophic no system, no institution, no mode of production could remain impervious. Especially transformed were Europe's labour markets. In England, those

who survived the Black Death lived to enjoy 'the sole golden age of the English peasantry' (Hatcher 1987, p. 281). The Statute of Artificers notwithstanding, the general response of the manorial economy to the desperate scarcity of labour, the abundance of abandoned land, and rising prices in grain markets was to relax feudal constraints on the mobility of labour, to raise nominal wages, lower rents, and make concessions on customary dues and obligations. In a word, peasants under villein tenure were able to secure copyhold and leasehold tenancies. By the 17th century, England had become 'a peasant-free zone' (R.M. Smith in Scott 1998, p. 346).

In Russia, in stunning contrast, the response to depopulation was the establishment of serfdom. In the face of acute labour scarcity and expanding grain markets, the Boyars demanded tighter and tighter legal restrictions on the mobility of peasants until, by an Act of the Duma in 1649, serfs and their households were made the personal property of the lord – effectively slaves (Domar 1970, pp. 13–32).

Thus, 'peasant' in the European context came to be a status that defies definition: some were well on the way to yeomanry by 1600, others were only a technicality away from a heritable condition of slavery.

Marketplaces and Peasants

From the 14th century to the 18th, the history of marketplaces is linked to the history of peasants, as that of peasants is to marketplaces. In Brueghel's exuberant paintings of peasants tumbling about in overflowing marketplaces the link has become an icon of Flemish art. 'Historically, peasants only exist when markets exist, even if they do not fully participate in them' (Scott 1998, p. 2).

Having for well over a century minutely observed a vast number of peasant societies, the characteristics of peasant markets, and the behaviour of peasants with respect to them, it has been anthropology, rather than economics, that has become, to use Grantham's phrase (1997, p. 19), 'the referent social science' of the peasant economy. Peasant villages, whether in Indo-China, West Africa, the Stone Age, or – improbably – colonial New England, are said to have this in common: that the village marketplace is 'embedded in' the social fabric of the community; the rules governing it, and the motivations of buyers and sellers and borrowers and lenders transacting in it are subordinate to and constrained by the non-market values of the ambient culture, and stand as epicentres of resistance to the encroachment of the 'disembedded', proto-capitalist, dynamic, hegemonic Market, with a capital M.

Influential as this model has been, there is a counter-narrative, accepted even by some anthropologists. Thus, for example, among the Panajachelenos of the Guatemalan highlands, 'Commerce is the breath of life' (Tax 1963, p. 132). According to Sol Tax, both men and women spend more time buying and selling and talking about buying and selling than doing anything else. When they have nothing to sell, they buy something in a cheap market and carry it to a dear one to sell. The Maoris of New Zealand, the Ifugao of the Philippines, the Senegambian, Afikpo, Esusu, Yoruba, Hausa, Tiv and Dahomean peoples of West Africa, the northeastern Malays, the Javanese and the Trobriand Islanders have all been found to engage in price- and quantity-bargaining, to 'seek maximal advantage' in marketplace transactions (Firth and Yamey 1964), and to exhibit in their market-dependence the same U-shaped relationship to their disposable assets as is associated with risk-averse behaviour in developed economies.

A case in point is a new study of African markets by Marcel Fafchamps (2004) who argues that sub-Saharan Africa today is decisively 'market-oriented'. The development of Africa, Fafchamps insists, has been impeded not by the cultural incongruity of markets, or the absence of markets, or the lack of a market *mentalité,* but by the accommodation that traders in sub-Saharan markets have had to make to the ubiquity of market imperfections. In the absence of well-defined property rights, intermediary institutions, notaries public and contract enforcement, the sole bar against asymmetric information, adverse

selection and moral hazard has been what might be called 'insider trading': repeated transactions among networks of friends and relations bound to one another by webs of trust. In the context of sub-Saharan Africa these webs, although intimate, are not cooperative, says Fafchamps, but strategic; the object is not to avoid the discipline of the market but more nearly to satisfy its assumptions.

Few peasant economies fit the 'moral economy' model (Rothenberg 1992, ch. 2). And those that do may, like 15th-century French villages, have been torn by strife, the collective experience 'rubbed raw' (Hoffman 1996, p. 77) by the face-to-face pursuit of what Avner Offer (1997) has called 'regard'. But the principal critique of a non-market model is that it is a steady-state model, and that has tended to impart a Durkheimian steady-state bias to peasant studies. It is a bias very much 'at home' with a methodology in anthropology itself in which scrutiny ever closer is leveled at subjects ever narrower – from tribe, to village, clan, household, extended family, nuclear family, gender – thereby surrendering to economists the trade *among* groups and the articulation *between* marketplaces that generates growth. It is in regulating that process that the community asserts its dominion over what Braudel (1982, p. 58) has called 'the insidious tentacles of the economy'.

Regulating the Marketplace

Such interventions have a very long history, coming down to us, Blackstone thought, from Saxon times when no title to goods valued above 20 pence could change hands without witnesses. But such impediments to trade could not have persisted were it not for law. It is the province of any law worthy of the name to recognize in 'You have got what belongs to me' its sphere of action (Pollock and Maitland, 1895, p. 33), and nowhere more urgently than at the point of transferring the ownership of chattels by sale. For what but law can distinguish sale from theft, right from use, ownership from possession, *dominium* from *seisin*?

As the frequency of trade increased in the 13th century, exchange in an open market established by royal grant – a 'market ouvert' – sufficed, in lieu of witnesses, to divest a seller. With outdoor marketplaces came rules, regulations and ordinances establishing, locating, restricting, and supervising open markets, covered markets, fairs, merchant courts and piepoudre courts for itinerant merchants; appointing ringers of opening and closing bells, monitors of weights and measures, collectors of fees and fines, inspectors of quality, enforcers of Just Prices, and wardens to patrol the perimeters of the marketplace. This regulatory 'apparatus' was brought to the American colonies and given the force of law in the municipal marketplaces of all large towns in Massachusetts and throughout the colonies.

'The Laws and Liberties of Massachusetts', codified between 1641 and 1691, declared the taking of excessive wages by mechanics and day labourers and the charging of unreasonable prices by shopkeepers and merchants punishable by double restitution or imprisonment. They set the weight of the pennyloaf of white bread, and authorized the selection of two able persons annually to enter into the houses of all bakers 'as oft as they see cause' to inspect and weigh all bread found there on pain of forfeiture. Two persons were appointed annually to ascertain the range of prevailing wheat prices each month and set the price at which bakers shall bake their bread. 'The Laws and Liberties' also set the days of the week when a marketplace shall be kept in Boston, Salem, Lynn and Charlestown, set the times of the ringing of the opening and closing bells, forbade all trade outside the perimeter of the marketplace, and set the two days a year when Boston, Salem, Watertown and Dorchester shall have fairs (Cushing 1976).

In 1737, Faneuil Hall, Boston's beleaguered marketplace, was besieged by farmers from the surrounding hinterland who 'donned the livery of heaven' (disguised themselves as clergy) and burned it to the ground (Brown 1900, ch. 8). By the early 1820s, the regulated 'market ouvert' and the legal doctrine of implied contract expressed by it were abandoned in favour of the rule of *caveat emptor*. Contract, not Community, would come increasingly to regulate markets.

Marshallian Markets: The Marketplace After 1900

With the publication of Alfred Marshall's *Principles of Economics* in 1900, economics became the referent social science of markets. It may therefore come as something of a surprise to discover that in this, the urtext of economics, marketplaces have vanished! Marshall's definition of a market is of an abstraction, outside of spatial coordinates, oblivious of cultural context, and functioning homeostatically in accordance with laws of its own making. 'The distinction of locality is not necessary,' he wrote. 'Economists understand by the term "Market" not any particular market-place in which things are bought and sold, but the whole of any region in which buyers and sellers are in such free intercourse with one another that the prices of the same good tend to equality easily and quickly' (Marshall 1890, p. 324).

Thus, the market is not a place but a *process* that expands in space and unfolds in time, driven by the pace at which different prices for the same good converge toward a single price. Called the Law of One Price, that convergence is the unintended consequence of arbitrage between buyers seeking cheap markets and sellers seeking dear markets.

But as long as a wedge between 'cheap' and 'dear' markets persists, that is, as long as the convergence process is incomplete, marketplaces remain significant, not as 'spots of ground' but as transitional nodes of price-formation that become 'folded in' as the market process advances along its dendritic expansion-path. The story we have followed for 60,000 years has not, even in a Marshallian sense, become irrelevant.

At the same time, economics itself, as the referent social science of markets, is expanding its narrow field of vision beyond its Marshallian boundaries in an attempt to comprehend the wider social and psychological foundations of economic behaviour. I am reminded of the earthquake of 8 October 2005 which struck high up in the Kashmiri Himalayas. Within a week of the catastrophe, survivors set up a village marketplace (BBC News). Seventy-three thousand had been killed, three million were made homeless, none had necessities, none had surpluses, but within days they made a marketplace.

See Also

- ▶ Braudel, Fernand (1902–1985)
- ▶ Domar, Evsey David (1914–1997)
- ▶ Market Institutions
- ▶ Peasant Economy
- ▶ Peasants
- ▶ Polanyi, Karl (1886–1964)

Bibliography

Allen, J.P. 2002. *The Heqanakht Papyri*. New York: Metropolitan Museum of Art and Yale University Press.

Braudel, F. 1982. *The wheels of commerce: Civilization & capitalism, 15th–18th century*, vol. 2. New York: Harper and Row.

Brown, A.E. 1900. *Faneuil Hall and the Faneuil Hall market, Or Peter Faneuil's gift*. Boston: Lee and Shepard.

Curtin, P.D. 1975. *Economic change in precolonial Africa: Senegambia in the era of the slave trade*. Madison: University of Wisconsin Press.

Cushing, D. 1976. *The laws and liberties of Massachusetts, 1641–1691*, 3 vols. Wilmington: Scholarly Resources.

Domar, E.D. 1970. The causes of slavery and serfdom: A hypothesis. *Journal of Economic History* 30: 18–32.

Ellis, F. 1988. *Peasant economics: Farm households and agrarian development*. Cambridge: Cambridge University Press.

Fafchamps, M. 2004. *Market institutions in Sub-Saharan Africa*. Cambridge, MA: MIT Press.

Firth, R., and B.S. Yamey, eds. 1964. *Capital, saving and credit in peasant societies: Studies from Asia, Oceania, the Caribbean, and Middle America*. Chicago: Aldine.

Gonzales, R. 2005. The geography of the Silk Road. http://www.humboldt.edu/~geog3091/ideas/raysilk.html. Accessed 21 July 2005.

Grantham, G. 1997. The shards of trade: archaeology and the economic history of the super-long-run. Paper delivered at the Economic History Association Conference, August 1997, p. 18.

Hatcher, J. 1987. English serfdom and villeinage: Towards a reassessment. In *Landlords, peasants and politics in medieval England*, ed. T. Aston. Cambridge: Cambridge University Press.

Hoffman, P.T. 1996. *Growth in a traditional society: The French countryside, 1450–1815*. Princeton: Princeton University Press.

Hughes, J.R.T. 1976. *Social control in the colonial economy*. Charlottesville: University Press of Virginia.

Kohn, M. 2003. Organized markets in pre-industrial Europe. Working paper. Dartmouth College, July 2003.
Marshall, A. 1890. *Principles of economics,* 8th edn. London: Macmillan.
Mazower, M. 2004. *Salonica, city of ghosts: Christians, Muslims, and Jews, 1430–1950*. New York: Alfred Knopf.
McCabe, I.B., G. Harlaftis, and I. Minoglou, eds. 2005. *Diaspora entrepreneurial networks: Four centuries of history*. New York: Berg.
McNeill, W.H. 1976. *Plagues and peoples*. Garden City: Anchor/Doubleday.
Offer, A. 1997. Between the gift and the market: The economy of regard. *Economic History Review* 50: 450–476.
Polanyi, K. 1944. *The great transformation: The political and economic origins of our time*. Boston: Beacon Press.
Pollock, F. and Maitland, F. 1895. *The history of English law before the time of Edward*, vol. 2. Cambridge: Cambridge University Press, 1968.
Rothenberg, W.B. 1992. *From market-places to a market economy: The transformation of rural Massachusetts, 1750–1850*. Chicago: University of Chicago Press.
Scott, T., ed. 1998. *The peasantries of Europe: From the fourteenth to the eighteenth centuries*. New York: Addison Wesley Longman.
Tax, S. 1963. *Penny capitalism: A Guatemalan Indian economy*. Chicago: University of Chicago Press.

Markets

Geoffrey M. Hodgson

Abstract

Until recently, and despite strong interest in market outcomes, economists have paid relatively little attention to the institutional structure of markets. This article considers the historical evolution of markets and poses some dilemmas concerning their definition. Several alternative definitions are considered, involving different degrees of historical specificity. It is argued that developments in economics and elsewhere since the 1980s point to a more nuanced view of markets, involving a recognition of different types of market mechanisms and institutions. These developments include work in experimental economics and auction theory. A definition of markets is offered that is consistent with these developments.

Keywords

Agent-based models; Amsterdam Bourse; Auctions; Auctions; Austrian economics; Bonded labour; Central planning; Clark, J. M; Coase, R; Commodity exchange; Commons, J; Contemporary capitalism; Contracts for service; Cultural norms; Dutch East India Company; Economic sociology; Experimental economics; Game theory; General equilibrium; German Historical School; Gift-exchange; Globalization; Hayek, F; Hobson, J; Imperfect information; Information; Labour market contracts; Law of one price; Learning; London Stock Exchange; Market institutions; Marketplaces; Markets; Marriage markets; Marxism; Mises, L. von; Missing markets; New institutional economics; New York Stock Exchange; North, D; Parsons, T; Polanyi, K; Property rights; Rau, K. H; Relational exchange; Robbins, L; Slavery; Smith, V; Stigler, G; Trust; Weber, M

JEL Classifications

D01

Markets dominate modern life, and economists have for long been concerned about market prices, but, despite this ongoing preoccupation, until recently there has been little discussion of the nature and operation of markets themselves.

No fewer than three Nobel Laureates in economics have noted this paradox. George Stigler (1967, p. 291) wrote: 'The efficacy of markets should be of great interest to the economist: Economic theory is concerned with markets much more than with factories or kitchens. It is, therefore, a source of embarrassment that so little attention has been paid to the theory of markets and that little chiefly to speculation.' Stigler made a plea for the theoretical study of markets, which for a long time went unheard.

Ten years later Douglass North (1977, p. 710) similarly remarked: 'It is a peculiar fact that the literature on economics and economic history contains so little discussion of the central institution that underlies neoclassical economics – the market'. Another 11 years had passed when

Ronald Coase (1988, p. 7) observed that 'in modern economic theory the market itself has an even more shadowy role than the firm'. Economists are interested only in 'the determination of market prices' whereas 'discussion of the market place itself has entirely disappeared'.

Economists have had little to say about the nature of markets, other than classifying them by their degrees of competition and their numbers of buyers and sellers. Beyond this, the institutional aspects of markets have been widely neglected. For much of the 20th century there has been little discussion of how specific markets are structured to select and authenticate information, and of how specific prices are actually formed. Furthermore, 'the market' was treated as a relatively homogeneous and undifferentiated entity, with little consideration of different market mechanisms and structures. When market mechanisms were addressed this was typically confined within the framework of general equilibrium theory, with relatively little attention to the institutional details and alternative market structures.

Inspection of standard economics textbooks confirms these observations. While market outcomes such as prices are always central to the discussion, there is generally little consideration of the detailed rules and mechanisms through which prices are formed, and the concept of the market itself often goes undefined. Indeed, there is an entry on markets in neither the massive 1968 edition of the *Encyclopaedia of the Social Sciences* nor the otherwise comprehensive 1987 edition of *The New Palgrave: A Dictionary of Economics.*

Three questions arise. First, what briefly is the nature of markets and how can the market be defined? Second, why has the specific anatomy of markets been neglected by economists? Third, what recent developments in economics and elsewhere help to remedy the deficiency? After a brief historical discussion this article addresses these questions.

Historical Background

Goods have changed hands within human societies for hundreds of thousands of years. However, much of this internal circulation was powered by custom and tradition. Transfers of goods often involved ceremony and personal, reciprocal actions. These personal and kin-based exchanges contrast with the organized and competitive pecuniary ambiance of modern markets. Ceremonial transfers involved 'the continuous definition, maintenance and fulfilment of mutual roles within an elaborate machinery of status and privilege' (Clarke 1987, p. 4). Most of this internal circulation of goods was devoid of any conception of the voluntary, contractual transfer of ownership or property rights. These reciprocal transfers of goods were more to do with the validation of custom and social rank.

Something more akin to trade existed at least as far back as the last ice age. However, this trade was largely peripheral and occurred at the meeting of different tribal groups. As Max Weber (1927, p. 195) attested, it did not take place 'between members of the same tribe or of the same community' but was 'in the oldest social communities an external phenomenon, being directed only towards foreign tribes'. This contention that trade began externally and between communities rather than within them has withstood subsequent scholarly examination. Trade was typically a collective and *inter-social* enterprise between one tribe and another.

With the rise of the ancient civilizations, both external and internal trade increased substantially. The development of money and coinage facilitated its expansion. A definable internal commodity market (or agora), with multiple buyers and sellers, first appeared in a designated open space in Athens in the sixth century BC (Polanyi 1971; North 1977). The *agora* opened frequently and had strict trading rules. At around the same time there existed an annual auction market on Babylonia: young women were put on display and male bidders competed for marriage rights (Cassady 1967). Nevertheless, some scholars have warned against the view that these ancient civilizations were generally and predominantly market economies (Finley 1962; Polanyi et al. 1957). By contrast, researchers such as Peter Temin (2006) have argued that the Roman Empire in particular contained developed and interlocking markets

with variable prices, albeit without a highly developed banking system and with a relatively limited market for capital.

European and Mediterranean trade contracted after the fall of the Roman Empire. When commerce began to develop again in medieval times, internal markets then had a limited role in the medieval economy. 'Strange though it may seem', wrote the historian Henri Pirenne (1937, p. 140), 'medieval commerce developed from the beginning not of local but of export trade.' Although there are likely to have been other earlier organized markets in England, systematic evidence of the king enforcing his right to license all markets and fairs does not appear until the 13th century.

Markets for slaves existed in classical antiquity and persisted in some regions until the modern era. By contrast, feudal serfs were not owned as chattels, but they did not enjoy the right to choose their masters. Feudal institutions, driven by traditional obligations rather than voluntary contract, meant that the hiring of labourers was marginalized and markets for wage labourers were rare. With the decline of bonded labour, which began as early as the 14th century in England, employment contracts were limited largely to casual labourers, alongside a large number of self-employed producers and others in peasant family units. In England it was not until about the 18th century that a class of potentially mobile wage labourers emerged who constituted the most important source of labour power. Organized markets for employees, involving labour exchanges or employment agents, did not become prominent until the 19th century.

To turn to capital markets, an early market for debts was the French *courratier de change* in the 12th century. In the 13th century, after the development of a banking system in Venice, trade began in government securities in several Italian cities. In 1309 a 'Beurse' was organised in Bruges in Flanders, named after the Van der Beurse family, who had previously hosted regular commodity exchanges in their residence. Soon after, similar 'Beurzen' opened in Ghent and Amsterdam. In 1602 the Dutch East India Company issued the first shares on the Amsterdam Bourse or Stock Exchange. The London Stock Exchange, founded in 1801, traces its origins to 1697 when commodity and stock prices began to be published in a London coffee house. The origins of the New York Stock Exchange go back to 1792, when 24 stockbrokers organized a regular market for stocks in Wall Street. Accordingly, developed capital markets first appeared in the Netherlands in the 17th century and later spread to other countries.

Overall, in the last 400 years markets have expanded enormously in scope, volume and economic importance. Markets have come to pervade internal as well as external trade and to dominate the global economic system. The modern era of globalization is often identified with the growth of global commodity and financial markets since the middle of the 19th century.

This brief historical sketch is background for the task of defining markets. At least three options emerge, involving different degrees of historical specificity. The broadest definition would be to use the term 'market' to refer to all forms of transfer of goods or services between persons, including the age-old customary or ceremonial transfers within tribes and households, exchanges of property between tribes, and modern organized markets with multiple buyers and sellers. We consider this option and some more restrictive alternatives next.

The Nature of Markets

The Austrian school economist Ludwig von Mises (1949) is exceptional among economists in devoting a lengthy chapter to 'the market'. He sees the market economy as 'the social system of the division of labour under private ownership of the means of production' (1949, p. 257). He explicitly excludes economies under social or state ownership of the means of production from this category, but nevertheless regards such systems as strictly 'not realizable'. Consequently, the historical and territorial boundaries of his concept of the market depend very much on what is regarded as 'private ownership'. He associates private ownership with the rise of civilization, and defines ownership in terms of full control of the services that derive from a good, rather than in legal terms. Together these specifications amount to a definition of the market that embraces all forms of trade or exchange that

involve private property, defined loosely as assets under private control.

Although von Mises associates secure private property and exchange with the rise of civilization, these terms are defined in a manner that does not exclude their application to earlier periods of human history. It then becomes problematic whether or not ceremonial transfers and ritualistic gift-giving are regarded as 'exchanges' of 'property' and whether or not these activities come within the sphere of 'the market'. Essentially, the historical compass of the latter term depends very much on what we mean by notions such as exchange and property.

In downplaying the legal aspects of property and exchange, von Mises also fails to probe the nature of the rights that form part of the exchange. Instead he sustains the notion that uncoerced and informed consent by the parties to the transaction is a sufficient basis to constitute the contractual and property rights involved. A problem with this idea is that mutual individual consent itself requires a legislative and institutional framework to legitimize, scrutinize and protect those individual rights. The importance of this legal and constitutional framework is widely recognized, including by other Austrian theorists such as Friedrich Hayek (1960). Several historical cases of the spontaneous evolution of systems of enforced property rights do exist, but they generally rely on reputational and other monitoring mechanisms that are more difficult to sustain in large-scale, complex societies (Sened 1997).

An alternative intellectual tradition places more emphasis on the legal and statutory basis of individual rights. This approach pervaded the 19th-century German Historical School and their predecessors such as Karl Heinrich Rau, and continued into the 20th century in the original American Institutionalist School, particularly in the writings of John Rogers Commons. Both Rau and Commons (1924) argued that exchange is more than a voluntary and reciprocal transfer of resources: it also involves the contractual interchange of statutory property rights. For them, exchange had to be understood and analysed in terms of the key institutions that are required to sustain it.

This narrower and more legalistic understanding of private property and contractual exchange confine them both in longevity and scale. Statutorily endorsed property rights, applied to moveable goods and services, were not codified until the ancient civilizations. In feudal times, much of the transfer of goods and services was achieved by custom or coercion rather than by contract and consent. Indeed, economic historians such as North (1981), who attempt to explore the origins of modern markets and commodity exchange, generally focus on the late medieval or early modern period as the era in which well-defined individual property rights began to spread widely from specific parts of the world.

A second important dilemma emerges. This is whether the market is regarded as coextensive with the exchange of private property per se or whether it is given an even narrower meaning and used to refer to forms of *organized* exchange activity. Two major factors lead us to consider an even narrower meaning for the term.

The first consideration is the commonplace use of the term 'market' itself and its equivalent in other languages. The word 'market' originally appeared as a noun to describe a specific place where people gathered and exchanges of a particular kind took place. The first market in Athens in the sixth century BC had rules concerning who could buy or sell, what could be bought or sold, and how trading should take place. In medieval England markets were permitted by royal charters and located in specific towns. In Europe and elsewhere in the last 300 years organized town and village markets have become commonplace. There are also permanent buildings that function as 'exchanges' for agricultural products, minerals, financial stocks, and so on. Although it has acquired additional meanings, the noun 'market' still refers to a place or gathering where trade is organized.

The second issue is the existence of a well-researched form of exchange that takes place in different contexts and involves different considerations. In three seminal and influential works, George B. Richardson (1972), Victor P. Goldberg (1980) and Ronald Dore (1983) point out that many real-world commercial

transactions do not take place in the competitive arena of a market. Instead they involve firms in ongoing contact, which exchange relevant information before, during and after the contract itself. The relationship is durable and the contract is often renewed. This is most often described as 'relational exchange'. A question in the derivative literature is to examine the reason for the mutual choice of an ongoing exchange relationship rather than the more competitive institution of the market. Among the explanations is the importance of establishing ongoing trust in circumstances of uncertainty where product characteristics are complex, relatively unique or involve continuous potential improvements. Whatever the reason for its existence, such relational contracting is very different from the more anonymous exchanges in organized markets. Relational exchanges are nevertheless still contractual exchanges of property rights, in their fullest and most meaningful sense. If they are distinguished by definition from market exchanges, then not all exchanges take place in markets. Furthermore, the exchange of goods or services that are strictly unique may be regarded as a non-market phenomenon, even if the exchange is not relational. The term 'market' is thus reserved for forms of exchange activity with many similar exchanges involving multiple buyers or sellers.

In part, it is the degree of organization of exchange activity that makes markets different from relational exchange. In financial markets, for example, there are typically strict rules concerning who can trade and how trading should be conducted. In such relatively volatile markets, specific institutions sift information and present it to traders to help the formation of price expectations and norms. Market institutions in other contexts monitor the quality of goods and the instruments of weight and measure. Within these structures, trading networks emerge on the basis of business connections and reputations.

Modern telecommunications mean that a market does not have to be organized in a specific location. Either bidders can communicate with the market centre over long distances, as with many financial markets, or the market *place* can itself disappear, as in the case of Internet-based markets, such as eBay. The latter case nevertheless remains a market, because it is an organized (virtual) forum, subject to specific procedures and rules.

We thus arrive at a definition of a market in the following terms. Markets involve multiple exchanges, multiple buyers or multiple sellers, and thereby a degree of competition. A market is defined as an institution through which multiple buyers or multiple sellers recurrently exchange a substantial number of similar commodities of a particular type. Exchanges themselves take place in a framework of law and contract enforceability. Markets involve legal and other rules that help to structure, organize and legitimize exchange transactions. They involve pricing and trading routines that help to establish a consensus over prices, and often help by communicating information regarding products, prices, quantities, potential buyers or possible sellers. Markets, in short, are organized and institutionalized recurrent exchange.

Of course, it is often difficult to draw the line between organized and relational exchange. There are many possible intermediate cases. However, such difficulties are typical when dealing with highly varied phenomena and are commonplace in some other sciences, notably biology. Similar difficulties exist in distinguishing other economic forms, such as making the important distinction between employment contracts and contracts for services. Nevertheless, such distinctions are important. The difficulty of defining a species does not mean that species should not be defined.

The operation of the law of one price is often taken as an indication of the existence of a market. Of course, imperfect information and quality variations can explain variations within a market from a single price. Nevertheless, the organized competition of the market and its associated information facilities are necessary institutional conditions for any gravitation by similar commodities to a single price level.

Taking stock, we may contrast the narrower definition of the market given above – as an institution with multiple buyers or multiple sellers, and recurrent exchanges of a specific type of commodity – with the much broader definitions raised earlier. These differences in definition do

not simply affect the degree of historical specificity of 'market' phenomena, they also sustain different theoretical frameworks and promote different questions for research. Some explanations for this divergence arise in the next section.

Why Have Economists Neglected the Institutional Character of Markets?

For much of the 20th century, the institutional character of markets has been neglected by economists because institutions generally have been neglected. The exceptions consist of economists who placed a special emphasis on institutions. The institutional character of markets was emphasised by German historical economists such as Gustav Schmoller and Werner Sombart in the 19th century (Hodgson 2001). The British dissident economist John A. Hobson (1902, p. 144) wrote: 'A market, however crudely formed, is a social institution'. Likewise, for the American institutionalist John Maurice Clark (1957, p. 53): 'the mechanism of the market, which dominates the values that purport to be economic, is not a mere mechanism for neutral recording of people's preferences, but a social institution with biases of its own'. Coase, North and others have effectively revived an interest in the institutional structure of markets that was eclipsed by developments in mainstream economics during much of the 20th century.

A further clue to help explain why generations of economists have neglected the institutional character of markets lies in the preceding section, where the problem of defining the boundaries of key concepts such as property, exchange and market was raised. Many economists have maintained that the principles of the subject should be as universal as possible – like physics – to the extent that substantial consideration of historically or nationally specific institutional structures is lost. The idea that economics should be defined as a general 'science of choice' (Robbins 1932) is part of this tradition. Consequently, terms such as property, exchange and market are given a wide meaning. Accordingly, many forms of human interaction have been regarded as 'exchange' and the summation of such 'exchanges' as 'markets'. In these terms, there is little difficulty in applying these concepts to many different types of system, from tribal societies through classical antiquity to the modern capitalist world.

Consequently, the idea of the market assumes a de-institutionalized form, as if it was the primeval and universal ether of all human interactions. Whenever people gather together in the name of self-interest, then a market somehow emerges in their midst. The market springs up simply as a result of these spontaneous interactions: it results neither from a protracted process of multiple institution-building nor from the full development of a historically specific commercial culture.

Incidentally, many sociologists have also assumed a de-institutionalized concept of the market. This is partly the result of the influence of a notion, promoted by Talcott Parsons and others, that sociology should also aspire to a high degree of historical generality. It is also a result of the influence of Marxism within sociology. Despite its emphasis on historical specificity, Marxism also treats markets as uniform entities, ultimately permeated by just one specific set of pecuniary imperatives and cultural norms.

From the 1940s to the 1970s, general equilibrium theory provided the framework in which economists attempted to understand the functioning of markets in wide-ranging terms. Even here, however, some significant attention had to be paid to institutional mechanisms and structures. Something special like the 'Walrasian auctioneer' had to be assumed in order to make the model work (Arrow and Hahn 1971). Some elemental institutional structures had to be brought in to make the model function in its own terms. The limits to this project of theoretical generalization became more apparent in the 1970s, when it was shown that few general conclusions could be derived. In particular, Hugo Sonnenschein (1972) and others demonstrated that within general equilibrium theory the aggregated excess demand functions can take almost any form (Rizvi 1994).

The existence of 'missing markets' always poses a problem for the general equilibrium approach: a complete set of markets for all present and future commodities in all possible states of the

world is typically assumed as a basis for general clearance in all markets. However, if market institutions are themselves scarce and costly to establish, then some may be missing for that reason. Furthermore, while capitalism has historically promoted market institutions, modern developed capitalism prohibits several types of market, such as markets for slaves, votes, drugs, or futures markets for labour. In so far as capitalism makes such prohibitions, 'missing markets' are inevitable within capitalism.

The technical problems exposed by Sonnenschein and others led economists to shift their attention away from general equilibrium theory. Instead, game theory became the cutting edge of theoretical analysis. By its nature, game theory tends to lead to less general propositions and points instead to more specific rules and institutions. As game theory became fashionable in the 1980s, it became a theoretical tool in the 'new institutionalist' revival in economic theory.

The Revival of the Notion of Markets as Institutions

At least three further developments helped to promote the study of markets as social institutions. First, the basic theory of auctions emerged in the 1970s and 1980s (McAfee and McMillan 1987; Wolfstetter 1996). It was assumed from the outset that participants in an exchange did not have complete information, and on this basis it was shown that choices concerning auction forms and rules could significantly affect market outcomes. These ideas assumed centre stage in the 1990s with the use by governments of auction mechanisms in electricity and telecommunications deregulation, most notably in the selling of the electromagnetic spectrum for telecommunications services, and subsequently with the growth of auctions on the Internet (McAfee and McMillan 1996).

A second and closely related development was the rise of experimental economics, which began to be recognized as an important subdiscipline in the 1980s. Modern experimental economists, in simulating markets in the laboratory, have found that they have had also to face the unavoidable problem of setting up its specific institutional structure. Simply calling it a market is not enough to provide the experimenter with the institutionally specific structures and procedural rules. As leading experimental economist Vernon Smith (1982, p. 923) wrote: 'it is not possible to design a laboratory resource allocation experiment without designing an institution in all its detail'. Work within experimental economics has underlined the importance of these specific rules, by showing that market outcomes are sometimes relatively insensitive to the information processing capacities of the agents involved, because particular constraints govern the results (Gode and Sunder 1993).

In reality, each particular market is entwined with other institutions and a particular social culture. Accordingly, there is not just one type of market but many different markets, each depending on its inherent routines, cultural norms and institutional make-up. Differentiating markets by market structure according to textbook typology – from perfect competition through oligopoly to monopoly – is not enough. Institutions, routines and culture have to be brought into the picture. Experimental economists have discovered an equivalent truth in laboratory settings, and have learned that experimental outcomes often depend on the tacit assumptions and cultural settings of participants. Different types of market institution are possible, involving different routines, pricing procedures, and so on. This has been acknowledged by a growing number of economists, as the notion of a single universal type of market has lost credibility (McMillan 2002).

Third, these theoretical developments were dramatized by events. Following the collapse of the Eastern bloc in 1989–91, a number of economists presumed that many markets would emerge spontaneously in the vacuum created by the breakdown of central planning. This view turned out to be mistaken, as capital and other markets were slow to develop and their growth was thwarted by the lack of an appropriate institutional infrastructrure. Several formerly planned economics slipped back into recession. Critics such

as Coase (1992, p. 718) drew attention to the necessary institutional foundations of the market system: 'The ex-communist countries are advised to move to a market economy . . . but without the appropriate institutions, no market of any significance is possible'.

While sociologists, like economists, had previously paid relatively little attention to market institutions, the revitalization of the sub-discipline of economic sociology led to a series of studies by sociologists of financial and other markets (Abolafia 1996; Baker 1984; Burt 1992; Fligstein 2001; Lie 1997; Swedberg 1994; White 1981; 1988). These works show how specific networks and social relationships between actors structure exchanges, and how cultural norms govern market operations and outcomes. Similar considerations have emerged in empirical and simulation work by economists that stresses the importance of learning and previous experience in trading partner selection and in the decision to accept a transaction (Kirman and Vignes 1991; Härdle and Kirman 1995).

Taken as a whole, these literatures testify to a much more nuanced conception of market phenomena. As a result of all these developments, the treatment by economists and others of markets began to change. Both economists and sociologists are now paying detailed attention to the nature of specific market rules and mechanisms. A milestone paper by Alvin Roth (2002) challenges the view of a single universal theory of market behaviour. While those economists who had paid attention to different market mechanisms had typically been preoccupied with a search for 'optimal' rules and institutional forms, gradually this has become a will- o'-the-wisp with the realization that typical assumptions in the emerging literature concerning cognitive and information impairments have made this search difficult or impossible (Lee 1998; Mirowski 2007).

Nevertheless, while the search for optimal institutional blueprints is intractable, these theoretical developments have begun to provide an analytical framework within which the limits and potentialities of different types of market mechanism can be appraised. An outcome is to abandon the former widespread notion – shared by all kinds of theorists from Marxists to the Austrian School – that 'the market' is a singular type of entity entirely understandable in terms of the same principles or laws. While Hayek and his followers should be given inspirational credit for their emphasis on the informational limitations inherent in all complex economic systems, they stressed that markets are the most effective processors of information while downplaying or ignoring the differences between various types of market.

In this context, markets reappear as varied and historically specific phenomena. The general equilibrium approach has been overshadowed by an array of theoretical and empirical methodologies, including game theory, agent-based modelling, laboratory experimentation and real-world observation.

Conclusions

A number of options for defining a market have been outlined here. The broadest option is to regard the market as the universal ether of human interaction, depending on little more than the division of labour. A second option is to regard the market as synonymous with commodity exchange, in which case it dates at least as far back to the dawn of civilization.

By contrast, several considerations militate in favour of a narrower definition, and recent developments in economic theory point in this direction. In the narrower sense, markets are organized exchange. Where they exist, markets help to structure, organize and legitimize numerous exchange transactions. Pricing and trading procedures within markets help to establish a consensus over prices, and communicate information regarding products, prices, quantities, potential buyers or possible sellers.

Variation in market rules and procedures means that markets differ substantially, especially when we consider markets in different cultures. The markets of 2000 years ago were very different from (say) the electronic financial markets of today. In the real world, and even in a single country, we may come across many different

examples of the market. The market itself is neither a natural datum nor an ubiquitous ether, but is itself a social institution, governed by sets of rules restricting some and legitimizing other behaviours. Furthermore, the market is necessarily entwined with other social institutions, such as in many cases the local or national state. It can emerge spontaneously, but it can also be promoted or guided by conscious design.

A clear implication of this argument is that the unnuanced but familiar pro- and anti-market policy stances are both insensitive to the possibility of different types of market institution. Instead of recognizing the important role of different possible cultures and trading customs, both the opponents and the advocates of the market have focused exclusively on its general features. Thus, for instance, Marxists have deduced that the mere existence of private property and markets will themselves encourage acquisitive, greedy behaviour, with no further reference in their analysis to the role of ideas and culture in helping to form the aspirations of social actors. This is the source of their 'agoraphobia', or fear of markets. Obversely, overenthusiastic advocates of the market claim that its benefits stem simply and unambiguously from the existence of private property and exchange, without regard to possible variations in detailed market mechanism or cultural context. As strange bedfellows, both Marxists and some market advocates have underestimated the degree to which all market economies are unavoidably made up of densely layered social institutions.

See Also

- ▶ Arbitrage
- ▶ Auctioneer
- ▶ Auctions (Applications)
- ▶ Austrian Economics: Recent Work
- ▶ Capitalism
- ▶ Competition
- ▶ Competition and Selection
- ▶ Competition, Austrian
- ▶ Competition, Classical
- ▶ Computing in Mechanism Design
- ▶ Contemporary capitalism
- ▶ Economic Sociology
- ▶ Econophysics
- ▶ Efficient Markets Hypothesis
- ▶ Electricity Markets
- ▶ Existence of General Equilibrium
- ▶ Experimental Economics
- ▶ General Equilibrium
- ▶ General Equilibrium (New Developments)
- ▶ Institutionalism, Old
- ▶ Labour Market Institutions
- ▶ Land Markets
- ▶ Market Institutions
- ▶ Market Microstructure
- ▶ Marketing Boards
- ▶ Marketplaces
- ▶ Marriage Markets
- ▶ Marx's Analysis of Capitalist Production
- ▶ Property rights
- ▶ Tâtonnement and Recontracting
- ▶ Two-sided Markets

Bibliography

Abolafia, M. 1996. *Making markets: Opportunism and restraint on Wall Street*. Cambridge, MA: Harvard University Press.

Arrow, K., and F. Hahn. 1971. *General competitive analysis*. Edinburgh: Oliver and Boyd.

Baker, W. 1984. The social structure of a national securities market. *American Journal of Sociology* 89: 775–811.

Burt, R. 1992. *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Cassady, R. 1967. *Auctions and auctioneering*. Berkeley/Los Angeles: University of California Press.

Clark, J. 1957. *Economic institutions and human welfare*. New York: Alfred Knopf.

Clarke, D. 1987. Trade and industry in barbarian Europe till Roman times. In *The Cambridge economic history of Europe, volume II: Trade and industry in the middle ages*, 2nd ed, ed. M. Postan and E. Miller. Cambridge: Cambridge University Press.

Coase, R. 1988. *The firm, the market, and the law*. Chicago: University of Chicago Press.

Coase, R. 1992. The institutional structure of production. *American Economic Review* 82: 713–719.

Commons, J. 1924. *Legal foundations of capitalism*. New York: Macmillan.

Dore, R. 1983. Goodwill and the spirit of market capitalism. *British Journal of Sociology* 34: 459–482.

Finley, M. (ed.). 1962. *Second international conference of economic history, volume. I: Trade and politics in the ancient World*. New York: Arno.

Fligstein, N. 2001. *The architecture of markets: An economic sociology of twenty-first century capitalist societies*. Princeton: Princeton University Press.
Gode, D., 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.
Goldberg, V. 1980. Relational exchange: Economics and complex contracts. *American Behavioral Scientist* 23: 337–352.
Härdle, W., and A. Kirman. 1995. Nonclassical demand: A model-free examination of price quantity relations in the Marseille fish market. *Journal of Econometrics* 67: 227–257.
Hayek, F. 1960. *The constitution of liberty*. London/Chicago: Routledge and Kegan Paul, and University of Chicago Press.
Hobson, J. 1902. *The social problem: Life and work*. London: James Nisbet.
Hodgson, G. 2001. *How economics forgot history: The problem of historical specificity in social science*. London/New York: Routledge.
Kirman, Alan P., and A. Vignes. 1991. Price dispersion: Theoretical considerations and empirical evidence from the Marseilles fish market. In *Issues in contemporary economics: Proceedings of the ninth world congress of the international economic association*, ed. K. Arrow. New York: New York University Press.
Lee, R. 1998. *What is an exchange? The automation, management, and regulation of financial markets*. Oxford: Oxford University Press.
Lie, J. 1997. Sociology of markets. *Annual Review of Sociology* 23: 341–360.
McAfee, R.P., and J. McMillan. 1987. Auctions and bidding. *Journal of Economic Literature* 25: 699–738.
McAfee, R., and J. McMillan. 1996. Analyzing the airwaves auction. *Journal of Economic Perspectives* 10(1): 159–175.
McMillan, J. 2002. *Reinventing the bazaar: A natural history of markets*. New York/London: Norton.
McMillan, J. 2003. Market design: The policy uses of theory. *American Economic Review: Papers and Proceedings* 93: 139–144.
Mirowski, P. 2007. Markets come to bits: Evolution, computation and markomata in economic science. *Journal of Economic Behavior and Organization* 63: 209–242.
North, D. 1977. Markets and other allocation systems in history: The challenge of Karl Polanyi. *Journal of European Economic History* 6: 703–716.
North, D. 1981. *Structure and change in economic history*. New York: Norton.
Pirenne, H. 1937. *Economic and social history of medieval Europe*. New York: Harcourt Brace.
Polanyi, K. 1971. *Primitive and modern economics: Essays of Karl Polanyi*. Boston: Beacon Press.
Polanyi, K., C. Arensberg, and H. Pearson (eds.). 1957. *Trade and market in the early empires*. Chicago: Henry Regnery.
Richardson, G. 1972. The organisation of industry. *Economic Journal* 82: 883–896.
Rizvi, S. 1994. The microfoundations project in general equilibrium theory. *Cambridge Journal of Economics* 18: 357–377.
Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.
Roth, A. 2002. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica* 70: 1341–1378.
Sened, I. 1997. *The political institution of private property*. Cambridge: Cambridge University Press.
Smith, V. 1982. Microeconomic systems as an experimental science. *American Economic Review* 72: 923–955.
Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40: 549–563.
Stigler, G. 1967. Imperfections in the capital market. *Journal of Political Economy* 75: 287–292.
Swedberg, R. 1994. Markets as social structures. In *Handbook of economic sociology*, ed. N. Smelser and R. Swedberg. Princeton: Princeton University Press.
Temin, P. 2006. The economy of the early Roman empire. *Journal of Economic Perspectives* 20(1): 133–151.
von Mises, L. 1949. *Human action: A treatise on economics*. London/New Haven: William Hodge/Yale University Press.
Weber, M. 1927. *General economic history*. London: Allen and Unwin.
White, H. 1981. Where do markets come from? *American Journal of Sociology* 87: 517–547.
White, H. 1988. Varieties of markets. In *Social structure: A network approach*, ed. B. Wellman and S. Berkowitz. Cambridge, MA: Harvard University Press.
Wolfstetter, E. 1996. Auctions: An introduction. *Journal of Economic Surveys* 10: 367–420.

Markov Chain Monte Carlo Methods

Siddhartha Chib

Abstract

MCMC methods, an important class of Monte Carlo methods, have played a major role in the growth of Bayesian statistics and econometrics. In an MCMC simulation, one samples a given distribution (say the posterior distribution in a Bayesian model) by simulating a suitably constructed Markov chain whose invariant distribution is the target distribution. The Metropolis–Hastings algorithm and its

special case, the Gibbs sampler, are two common ways of devising an MCMC simulation. We discuss how these methods originate, discuss implementation issues and provide examples. The use of MCMC methods in Bayesian prediction and model choice problems is also discussed.

Keywords

Autocorrelation; Bayesian econometrics; Bayesian prior–posterior analysis; Bayesian statistics; Invariance; Latent variables; Marginal likelihood; Markov chain Monte Carlo methods; Model choice; Prediction; Proposal densities; Reversibility; Transition density

JEL Classifications

C10

Introduction

Markov chain Monte Carlo methods, popularly called MCMC methods, are a class of Monte Carlo methods for sampling a given univariate or multivariate probability distribution (the target distribution). These methods play a central role in the theory and practice of modern Bayesian methods where they are used for the numerical calculation of quantities (such as the moments and quantiles of posterior and predictive densities) that arise in the Bayesian prior–posterior analysis. They have transformed the fields of Bayesian statistics and econometrics.

Suppose that in a given Bayesian model the prior density is $\pi(\boldsymbol{\theta})$ and the sampling density or likelihood function is $f(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y}$ is a vector of observations and $\boldsymbol{\theta} \in \mathcal{R}^d$ is an unknown parameter. In the Bayesian context, inferences about $\boldsymbol{\theta}$ are based on the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})$. Now suppose that one is interested in finding the mean of the posterior density

$$E(\boldsymbol{\theta}|\,\mathbf{y}) = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

but that the integral cannot be computed analytically. In that case one can compute the integral by Monte Carlo sampling methods. The general idea is to calculate the integral from a sample

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \sim \pi(\boldsymbol{\theta}|\mathbf{y}),$$

that is drawn from the posterior density. This sample can be used to estimate the posterior mean and other features of the posterior density. For instance, the posterior mean can be estimated by the average of the sampled draws, and the quantiles of the posterior density by the quantiles of the sampled output.

The requisite sampling of the target density is made possible by MCMC methods. In a MCMC simulation, one samples the target density in an indirect way: by simulating a suitably constructed Markov chain whose invariant distribution is the target density. Then the draws beyond some chosen burn-in period are taken as a (correlated) sample from the target density. The defining feature of Markov chains is the property that the conditional density of $\boldsymbol{\theta}^{(j)}$ (the jth element of the sequence) conditioned on the entire preceding history of the chain depends only on the previous value $\boldsymbol{\theta}^{(j-1)}$. Denote this conditional density, the transition density of the Markov chain, by $p(\boldsymbol{\theta}^{(j-1)}, \cdot|\mathbf{y})$. Then, in the MCMC framework, a sample is produced by simulating the transition density as

$$\boldsymbol{\theta}^{(1)} \sim p\left(\boldsymbol{\theta}^{(0)}, \cdot\,|\mathbf{y}\right) \vdots\, \boldsymbol{\theta}^{(j)} \sim p\left(\boldsymbol{\theta}^{(j-1)}, \cdot\,|\mathbf{y}\right) \vdots$$

If we let the first n_0 cycles represent the burn-in phase, for some choice of n_0, the draws

$$\boldsymbol{\theta}^{(n_0+1)}, \boldsymbol{\theta}^{(n_0+2)}, \ldots, \boldsymbol{\theta}^{(n_0+M)}$$

are treated as those from $\pi(\boldsymbol{\theta}|\mathbf{y})$. Even though the sampled variates are correlated, laws of large numbers for Markov sequences can be used to show that, under regularity conditions, the sample average of any integrable function $g(\boldsymbol{\theta})$ converges to its posterior expectation:

$$M^{-1}\sum_{j=1}^{M} g\left(\boldsymbol{\theta}^{(j)}\right) \to \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \qquad (1)$$

as M becomes large.

There are two common ways of constructing a transition density $p(\boldsymbol{\theta}^{(j-1)},\cdot|\mathbf{y})$ whose limiting distribution is the required target density. One way is by a method called the Metropolis–Hastings (M–H) algorithm, which was introduced by Metropolis et al. (1953) and Hastings (1970). Key references about this method are Tierney (1994), and Chib and Greenberg (1995). A second approach is by the so-called Gibbs sampling algorithm. This method was introduced by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990), and was the impetus for the current interest in Markov chain sampling methods. A summary of many aspects of MCMC methods is contained in Chib (2001) while textbook accounts include Gilks, Richardson and Spiegelhalter (1996), Chen, Shao and Ibrahim (2000), Liu (2001) and Robert and Casella (2004).

Metropolis–Hastings Algorithm

Suppose that we are interested in sampling the target density $\pi(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta}$ is a vector-valued parameter and $\pi(\boldsymbol{\theta}|\mathbf{y})$ is a continuous density. The idea behind the M–H algorithm is to simulate a proposal value $\boldsymbol{\theta}'$ from a transition density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ that is convenient to stimulate but does not necessarily have the correct limiting distribution and then to subject the proposal value to a specific randomization to ensure that the resulting Markov chain has the correct limiting distribution.

To define the M–H algorithm, let $\boldsymbol{\theta}^{(j-1)}$ be the current value. Then the next value $\boldsymbol{\theta}^{(j)}$ is produced by a two-step process consisting of a 'proposal step' and a 'move step'.

- *Proposal step*: Sample a proposal value $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}|\mathbf{y})$ and calculate the quantity

$$\alpha\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}' \,|\, \mathbf{y}\right) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}'|\,\mathbf{y})}{\pi\left(\boldsymbol{\theta}^{(j-1)}|\,\mathbf{y}\right)} \frac{q\left(\boldsymbol{\theta}', \boldsymbol{\theta}^{(j-1)}|\,\mathbf{y}\right)}{q\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\,\mathbf{y}\right)}\right\}. \quad (2)$$

- *Move step*: Let $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}'$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$; remain at the current value $\boldsymbol{\theta}^{(j-1)}$ with probability $1 - \alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$.

In terms of nomenclature, the source density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ is called the candidate generating density or proposal density, and $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$ the *acceptance probability* or, more descriptively, the *probability of move*. Note also that the function $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}'|\mathbf{y})$ in this algorithm can be computed without knowledge of the norming constant of the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. In addition, if the proposal density is symmetric, satisfying the condition $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$, then the acceptance probability reduces to $\pi(\boldsymbol{\theta}'|\mathbf{y})/\pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$; hence, if $\pi(\boldsymbol{\theta}') \geq \pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$, the chain moves to $\boldsymbol{\theta}'$, otherwise it moves to $\boldsymbol{\theta}'$ with probability given by $\pi(\boldsymbol{\theta}'|\mathbf{y})/\pi(\boldsymbol{\theta}^{(j-1)}|\mathbf{y})$. The latter is the algorithm of Metropolis et al. (1953).

Remark 1: Derivation of the M–H algorithm A question of some interest is the justification of this two-step approach. This question was tackled by Chib and Greenberg (1995) who derived the method from the logic of reversibility. A Markov transition density $p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ is said to be reversible for $\pi(\boldsymbol{\theta}|\mathbf{y})$ if the following condition holds for every $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in the support of the target distribution:

$$\pi(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \pi(\boldsymbol{\theta}'|\mathbf{y})p(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}). \quad (3)$$

The reversibility condition is important because reversible chains are invariant. Invariance refers to the property that

$$\pi(\boldsymbol{\theta}'|\mathbf{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \quad (4)$$

which means that, if the transition density is invariant for the target density, then, once convergence is achieved, a subsequent value $\boldsymbol{\theta}'$ drawn from the transition density is also from the target density. To see that reversibility implies invariance one simply integrates both sides of Eq. (3) over $\boldsymbol{\theta}$. This leads to the invariance condition since $\int p(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\, d\boldsymbol{\theta}' = 1$ by virtue of being a transition density. Now consider the Markov chain induced by the

proposal density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$. Because this was formulated without the reversibility condition in mind it is unlikely to satisfy reversibility. Suppose that for a pair of points $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ it is true that

$$\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) > \pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}), \qquad (5)$$

which means informally that the process moves from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ too frequently and too rarely in the reverse direction. This situation can be corrected by reducing the flow from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ by introducing probabilities $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ and $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ of making the moves in either direction so that

$$\begin{aligned}&\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\\ &\quad = \pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y}).\end{aligned}$$

One now sets $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ to be as high as possible, namely, equal to 1. Solving for $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ one then gets

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})} \frac{q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})}.$$

Because one started from Eq. (5) this is clearly less than 1. On the other hand, if the inequality in Eq. (5) were reversed, the same argumentation leads to the conclusion that $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = 1$. Thus, on combining these two cases we reproduce the expression of $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ given in Eq. (2).

Remark 2: Transition density of the M–H chain The transition density of the M–H chain has two components – one for the move away from $\boldsymbol{\theta}$ and given by $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ and one for the probability of staying at $\boldsymbol{\theta}$ given by $r(\boldsymbol{\theta}|\mathbf{y}) = 1 - \int \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\, d\boldsymbol{\theta}'$. In particular,

$$\begin{aligned}p_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = {} & \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})\\ & + \delta_\theta(\boldsymbol{\theta}')r(\boldsymbol{\theta}|\mathbf{y})\end{aligned}$$

where $\delta_\theta(\boldsymbol{\theta}')$ is the Dirac-function at $\boldsymbol{\theta}$ defined as $\delta_\theta(\boldsymbol{\theta}') = 0$ for $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ and $\int \delta_\theta(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' = 1$. It is easy to check that the integral of the transition density over all possible values of $\boldsymbol{\theta}$ is 1, as required.

Remark 3: Convergence properties The theoretical properties of the M–H algorithm (in particular the ergodic behaviour of the chain from an arbitrarily specified initial value) depend crucially on the nature of the proposal density. One requirement is that the proposal density be everywhere positive in the support of the posterior density, which means that the M–H chain can make a transition to any point in its support in one step. Further discussion of the conditions is given in Tierney (1994) and Robert and Casella (2004).

Remark 4: Mixing The sampled values from the M–H algorithm (as from any Markov chain) are correlated. The goal in any particular application is to ensure that the serial correlation is not excessive. One diagnostic to check for the degree of serial correlation in the sampled draws is the *autocorrelation time* or *inefficiency factor* of each component $\boldsymbol{\theta}_k$ of $\boldsymbol{\theta}$ defined as

$$a_k = \left\{1 + 2\sum_{s=1}^{M}\left(1 - \frac{s}{M}\right)\rho_{k_s}\right\},$$

where ρ_{ks} is the sample autocorrelation at lag s from the M sampled draws $\theta_k^{(n_0+1)}, \ldots, \theta_k^{(n_0+M)}$. One can interpret this quantity in terms of the *effective sample size*, or ESS, defined for the kth component of $\boldsymbol{\theta}$ as $ESS_k = \frac{M}{a_k}$. With independent sampling the autocorrelation times are theoretically equal to 1, and the effective sample size is M. When the inefficiency factors are high, the effective sample size is much smaller than M.

Choice of Proposal Density

One family of candidate-generating densities is given by $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}' - \boldsymbol{\theta})$. The candidate $\boldsymbol{\theta}'$ is thus drawn according to the process $\boldsymbol{\theta}' = \boldsymbol{\theta} + \mathbf{z}$, where $\mathbf{z}$ follows the distribution q, and is called the *random walk M–H* chain. The random walk M–H chain is quite popular in applications. One has to be careful in setting the variance of $\mathbf{z}$ because if it is too large the chain may remain stuck at a particular value for many iterations, while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution.

Another possibility is to let $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = q(\boldsymbol{\theta}'|\mathbf{y})$, an *independence M–H chain* in the terminology of Tierney (1994). One way to implement such chains is by tailoring the proposal density to the target at the mode by a multi-variate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode (Chib and Greenberg 1994, 1995).

Yet another way to generate proposal values is through a Markov chain version of the accept–reject method (Tierney 1994; Chib and Greenberg 1995). To explain this method, suppose $c > 0$ is a known constant and $h(\boldsymbol{\theta})$ a source density. Let $C = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\mathbf{y}) \leq ch(\boldsymbol{\theta})\}$ denote the set of value for which $ch(\boldsymbol{\theta})$ dominates the target density. Given $\boldsymbol{\theta}^{(j-1)} = \boldsymbol{\theta}$ the next value $\boldsymbol{\theta}^{(j)}$ is obtained as follows. First, a candidate value $\boldsymbol{\theta}'$ is obtained, independent of the current value $\boldsymbol{\theta}$, by applying the accept–reject algorithm with $ch(\cdot)$ as the 'pseudo-dominating' density. The candidates $\boldsymbol{\theta}'$ that are produced under this scheme have density $q(\boldsymbol{\theta}'|\mathbf{y}) \propto \min\{\pi(\boldsymbol{\theta}'|\mathbf{y}); ch(\boldsymbol{\theta}')\}$. Then, the M–H probability of move is given by

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{\theta} \in C \\ 1/w(\boldsymbol{\theta}) & \text{if} \quad \boldsymbol{\theta} \notin C, \boldsymbol{\theta}' \in C \\ \min\{w(\boldsymbol{\theta}')/w(\boldsymbol{\theta}), 1\} & \text{if} \quad \boldsymbol{\theta} \notin C, \boldsymbol{\theta}' \notin C \end{cases} \tag{6}$$

where $w(\boldsymbol{\theta}) = c^{-1}\pi(\boldsymbol{\theta}|\mathbf{y})/h(\boldsymbol{\theta})$.

Example

To illustrate the M–H algorithm, consider the binary response data in Table 1, on the occurrence or non-occurrence of infection following birth by Caesarean section. The response variable y is 1 if the Caesarean birth resulted in an infection, and zero if not. There are three covariates: x_1, an indicator of whether the caesarean was non-planned; x_2, an indicator of whether risk factors were present at the time of birth; and x_3, an indicator of whether antibiotics were given as a prophylaxis. The data in the table contains information from 251 births. Under the column of the response, an entry such as 11/87 means that there were 98 deliveries with covariates (1,1,1) of whom 11 developed an infection and 87 did not. Suppose that the probability of infection for the ith birth ($i \leq 251$) is

Markov Chain Monte Carlo Methods, Table 1 Caesarean infection data

y(1/0)	$x1$	$x2$	$x3$
11/87	1	1	1
1/17	0	1	1
0/2	0	0	1
23/3	1	1	0
28/30	0	1	0
0/9	1	0	0
8/32	0	0	0

Source: Fahrmeir and Tutz (1994)

$$\Pr(y_i = 1|\mathbf{x}_i, \beta) = \Phi(\mathbf{x}_i'\beta), \tag{7}$$

$$\beta \sim N_4(0, 5\mathbf{I}_4) \tag{8}$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ is the covariate vector, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ is the vector of unknown coefficients, Φ is the cdf of the standard normal random variable and $\mathbf{I}_4$ is the four-dimensional identity matrix. The target posterior density, under the assumption that the outcomes $\mathbf{y} = (y_1, y_2, ..., y_{251})$ are conditionally independent, is

$$\pi(\beta|\mathbf{y}) \propto \pi(\beta)\prod_{i=1}^{251}\Phi(\mathbf{x}_i'\beta)^{y_i}\{1 - \Phi(\mathbf{x}_i'\beta)\}^{(1-y_i)}$$

where $\pi(\beta)$ is the density of the $N(0,10\mathbf{I}_4)$ distribution.

To define the Chib and Greenberg (1994) tailored proposal density, let

$$\hat{\beta} = (-1.093022\ 0.607643\ 1.197543 - 1.904739)'$$

be the maximum likelihood estimate and let

$$\mathbf{v} = \begin{pmatrix} 0.040745 & -0.007038 & -0.039399 & 0.004829 \\ & 0.073101 & -0.006940 & -0.050162 \\ & & 0.062292 & -0.016803 \\ & & & 0.080788 \end{pmatrix}$$

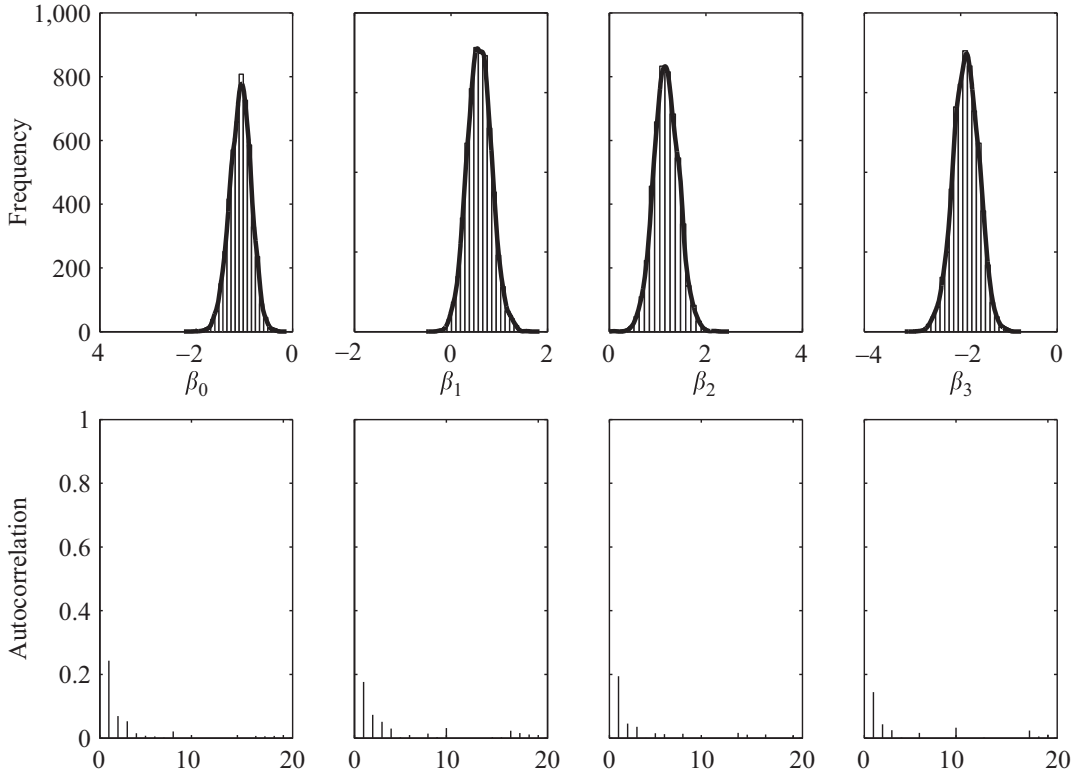
M

be the symmetric matrix obtained by inverting the negative of the Hessian matrix (the matrix of second derivatives) of the log-likelihood function evaluated at $\hat{\beta}$. To generate proposal values, we use a multivariate-t density with 15 degrees of freedom, location given by $\hat{\beta}$ and dispersion given by $\mathbf{V}$. The M–H algorithm is run for 5000 iterations beyond a burn-in of 100 iterations. The prior–posterior summary is reported in Table 2. It contains the first two moments (the mean and the standard deviation) of the prior and posterior and the 2.5th (lower) and 97.5th (upper) percentiles of the marginal densities of β.

In addition, we plot in Fig. 1 the four marginal posterior densities. These are derived by smoothing the histogram of the simulated values with a Gaussian kernel. In the same plot we also report the autocorrelation functions (correlation against

Markov Chain Monte Carlo Methods, Table 2 Caesarean data: prior–posterior summary based on 5000 draws (beyond a burn-in of 100 cycles) from the tailored M–H algorithm

	Prior		Posterior			
	Mean	Std dev	Mean	Std dev	Lower	Upper
β_0	0.000	2.236	− 1.080	0.220	− 1.526	− 0.670
β_1	0.000	2.236	0.593	0.249	0.116	1.095
β_2	0.000	2.236	1.181	0.254	0.680	1.694
β_3	0.000	2.236	− 1.889	0.266	− 2.421	− 1.385



Markov Chain Monte Carlo Methods, Fig. 1 Caesarean data with tailored M–H algorithm: marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

lag) for each of the sampled parameter values. The serial correlations decline quickly to zero indicating that the algorithm is mixing well.

Multiple-Block M–H Algorithm

When the dimension of $\boldsymbol{\theta}$ is large it is often necessary to divide the parameters into smaller groups or blocks and then to sample the blocks in turn. For simplicity suppose that two blocks are adequate and that $\boldsymbol{\theta}$ is written as $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_k \in \Omega_k \subseteq \mathcal{R}^{d_k}$. To sample these blocks let

$$q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2); q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1),$$

denote the two proposal densities, one for each block $\boldsymbol{\theta}_k$, where the proposal density q_k may depend on the current value of the remaining block. Also, define

$$\alpha(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) = \min\left\{\frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}) q_1(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2)}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2)}, 1\right\}$$

and

$$\alpha(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1) = \min\left\{\frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}) q_2(\boldsymbol{\theta}_2', \boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1)}, 1\right\}$$

as the probability of move for block $\boldsymbol{\theta}_k$ conditioned on the other block. Then, in what may be called the multiple-block M–H algorithm, one updates each block using an M-H step with the above probability of move, given the most current value of the other block. The method can be extended to several blocks in the same way.

Remark 5 An important special case arises if each proposal density is the full conditional density of that block. Specifically, if we set

$$q_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}),$$
$$q_1(\boldsymbol{\theta}_1', \boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$$

and

$$q_2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2' | \mathbf{y}, \boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}),$$
$$q_2(\boldsymbol{\theta}_2', \boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1) \propto \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})$$

then an interesting simplification occurs. The probability of move (for the first block) becomes

$$\alpha_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1' | \mathbf{y}, \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y}) \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y})}{\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{y}) \pi(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2 | \mathbf{y})}\right\} = 1,$$

and similarly for the second block, implying that, if proposal values are drawn from their full conditional densities, then the proposal values are accepted with probability one. This special case is called the Gibbs sampling algorithm.

The Gibbs Sampling Algorithm

The Gibbs sampling was introduced by Geman and Geman (1984) in the context of image processing and then discussed in the context of missing data problems by Tanner and Wong (1987). It was brought into prominence by Gelfand and Smith (1990) who demonstrated its use in a range of Bayesian problems.

The Algorithm

Suppose that the parameters are grouped into two p blocks $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)$ with the associated set of full conditional distributions

$$\{\pi(\boldsymbol{\theta}_1 | \mathbf{y}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p); \pi(\boldsymbol{\theta}_2 | \mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_p); \ldots \pi(\boldsymbol{\theta}_p | \mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d-1})\},$$

where each full conditional distribution is proportional to $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p) | \mathbf{y})$. Then, one cycle of the Gibbs sampling algorithm is completed by simulating $\{\boldsymbol{\theta}_k\}_{k=1}^{p}$ from these distributions, recursively refreshing the conditioning variables.

Sufficient Conditions for Convergence

Under rather general conditions, the Markov chain generated by the Gibbs sampling algorithm converges to the target density as the number of

M

iterations become large. Formally, if we let $p_G(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ represent the transition density of the Gibbs algorithm and let $p_G^{(M)}(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})$ be the density of the draw $\boldsymbol{\theta}'$ after M iterations given the starting value $\boldsymbol{\theta}_0$, then

$$\left\| p_G^{(M)}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}' | \mathbf{y} - \pi(\boldsymbol{\theta}' | \mathbf{y}) \right\| \to 0, \quad \text{as} \quad M \to \infty. \tag{9}$$

Roberts and Smith (1994) (see also Chan 1993) have shown that this convergence occurs under the following weak conditions: (i) $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$ implies there exists an open neighbourhood N_θ containing $\boldsymbol{\theta}$ and $\varepsilon > 0$ such that, for all $\boldsymbol{\theta}' \in N_\theta$, $\pi(\boldsymbol{\theta}'|\mathbf{y}) \geq \varepsilon > 0$; (ii) $\int \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_k$ is locally bounded for all k, where $\boldsymbol{\theta}_k$ is the kth block of parameters; and (iii) the support of $\boldsymbol{\theta}$ is arc connected.

MCMC Sampling with Latent Variables

MCMC sampling can involve not just parameters but also latent variables. This idea was called data augmentation by Tanner and Wong (1987) in the context of missing data problems.

To fix notations, suppose that $\mathbf{z}$ denotes a vector of latent variables and let the modified target distribution be $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. If the latent variables are tactically introduced, the conditional distribution of $\boldsymbol{\theta}$ (or sub-components of $\boldsymbol{\theta}$) given $\mathbf{z}$ may be easy to derive. Then, a multiple-block M–H simulation is conducted with the blocks $\boldsymbol{\theta}$ and $\mathbf{z}$ leading to the sample

$$\left(\boldsymbol{\theta}^{(n_0+1)}, \mathbf{z}^{(n_0+1)}\right), \ldots, \left(\boldsymbol{\theta}^{(n_0+M)}, \mathbf{z}^{(n_0+M)}\right) \sim \pi(\boldsymbol{\theta}, \mathbf{z} | \mathbf{y}),$$

where the draws on $\boldsymbol{\theta}$, ignoring those on the latent data, are from $\pi(\boldsymbol{\theta}|\mathbf{y})$, as required.

To demonstrate this technique in action, consider the probit regression example discussed in section "Example". Albert and Chib (1993) introduced a technique for this and related models that capitalizes on the simplifications afforded by introducing latent data into the sampling. The Albert–Chib method has found wide use and has made possible the routine analysis of models for categorical responses. To begin, let

$$z_i | \beta \sim N(\mathbf{x}_i' \beta, 1),$$
$$y_i = I[z_i > 0], \quad i \leq n,$$
$$\beta \sim N_k(\beta_0, \mathbf{B}_0). \tag{10}$$

This specification is equivalent to the probit model since $\Pr(y_i = 1 | \mathbf{x}_i, \beta) = \Pr(z_i > 0 | \mathbf{x}_i, \beta) = \Phi(\mathbf{x}_i' \beta)$. Now the MCMC sampling is based on the full conditional distributions

$$\beta | \mathbf{y}, \{z_i\}; \quad \{z_i\} | \mathbf{y}, \beta,$$

which are both tractable. In particular, the distribution of β conditioned on the latent data becomes independent of the observed data and has the same form as in the Gaussian linear regression model with the response data given by $\{z_i\}$ and is multivariate normal with mean $\hat{\beta} = \mathbf{B}\left(\mathbf{B}_0^{-1}\beta_0 + \sum_{i=1}^n \mathbf{x}_i z_i\right)$ and variance matrix $\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1}$. Next, the distribution of the latent data conditioned on the data and the parameters factor into a set of n independent distributions with each depending on the data through y_i:

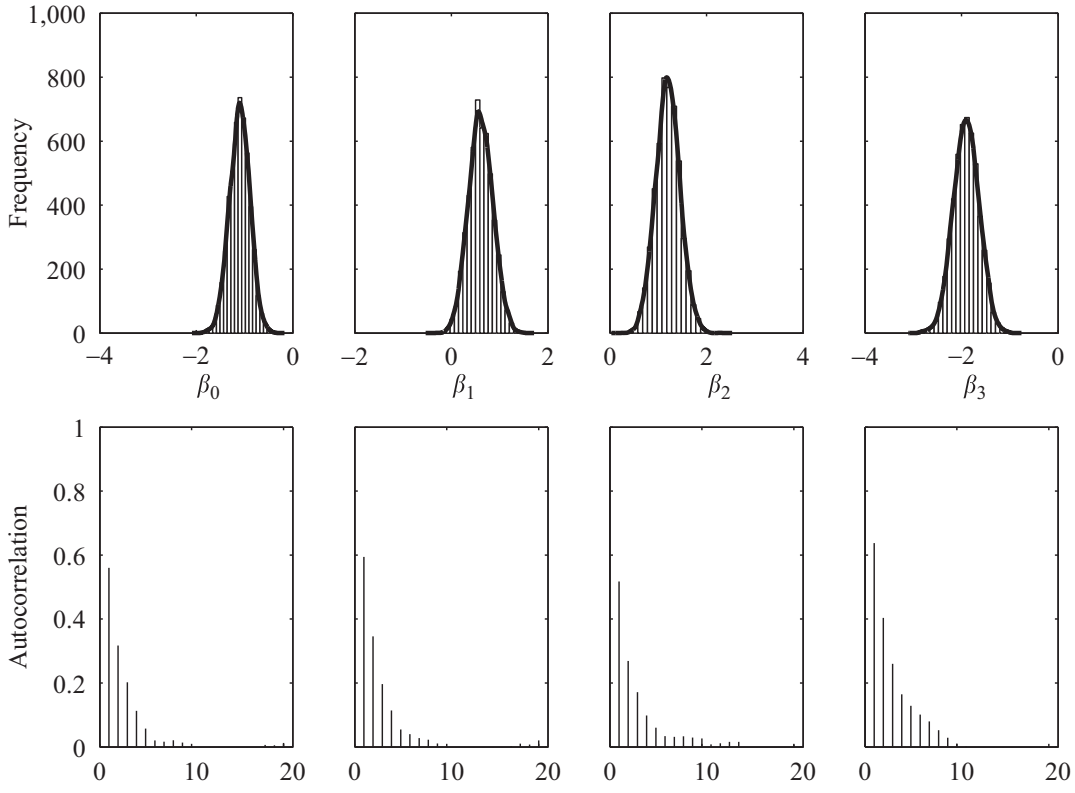
$$\{z_i\} | \mathbf{y}, \beta \overset{d}{=} \prod_{i=1}^n z_i | y_i, \beta,$$

where the distribution $z_i|\mathbf{y}_i, \beta$ is the normal distribution $z_i|\beta$ truncated by the knowledge of y_i; if $y_i = 0$, then $z_i \leq 0$ and if $y_i = 1$, then $z_i > 0$. Thus, one samples z_i from $TN_{(-\infty,0)}(\mathbf{x}_i'\beta, 1)$ if $y_i = 0$ and from $TN_{(0,\infty)}(\mathbf{x}_i'\beta, 1)$ if $y_i = 1$, where $TN_{(a;b)}(\mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ distribution truncated to the region (a, b).

We apply this method to the example considered in section "Example" above and report the results in Fig. 2. We see the close agreement between the two sets of results.

Strategies for Improving Mixing

In practice, while implementing MCMC methods it is important to construct samplers that mix well, where mixing is measured by the autocorrelation



Markov Chain Monte Carlo Methods, Fig. 2 Caesarean data with Albert–Chib algorithm: marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

time, because such samplers can be expected to converge more quickly to the invariant distribution.

Choice of Blocking

As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution.

Blocks can be combined by the method of composition. For example, suppose that $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ denote three blocks and that the distribution $\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_3$ is tractable (that is, can be sampled directly). Then, the blocks $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ can be collapsed by first sampling, $\boldsymbol{\theta}_1$ from $\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_3$ followed by $\boldsymbol{\theta}_2$ from $\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1; \boldsymbol{\theta}_3$. This amounts to a two-block MCMC algorithm. In addition, if it is possible to sample $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ marginalized over $\boldsymbol{\theta}_3$ then the number of blocks is reduced to one. Liu (1994) discusses the value of these strategies in the context of a three-block Gibbs MCMC chain. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains used to sample multivariate normal target distributions.

Tuning the Proposal Density

The proposal density in an M–H algorithm has an important bearing on the mixing of the MCMC chain. Chib and Greenberg (1994, 1995), Tierney (1994), Tierney and Mira (1999) and Liu (2001) discuss various possibilities for formulating proposal density that can be helpful in a variety of problems.

Prediction and Model Choice

In some settings, for example in models for time series data, an important goal is prediction. In the

Bayesian context, a future observation y_f is predicted through the (predictive) density defined as

$$f\left(y_f|\mathbf{y}\right) = \int f(y_f|\mathbf{y},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\,\mathbf{y})\,d\boldsymbol{\theta},$$

where $f\left(y_f|\,\mathbf{y},\mathcal{M},\,\boldsymbol{\theta}\right)$ is the conditional density of y_f given (y, $\boldsymbol{\theta}$). In general, the predictive density is not available in closed form. It can be shown, however, that, if one simulates $y_f^{(j)} \sim f\left(y_f|\mathbf{y},\boldsymbol{\theta}^{(j)}\right)$ for each sampled draw $\boldsymbol{\theta}^{(j)}$ from the MCMC simulation, then the collection of simulated values $\left\{y_f^{(1)},\ldots,y_f^{(M)}\right\}$ is a sample from $f(y_f|\mathbf{y})$. This simulated sample can be summarized in the usual way.

MCMC methods have also been widely applied to the problem of the model choice. Suppose that there are K possible models $\mathcal{M}_1,\ldots,\mathcal{M}_K$ for the observed data defined by the sampling densities $\{f(\mathbf{y}|\,\boldsymbol{\theta}_k,\mathcal{M}_k)\}$ and proper prior densities $\{p(\boldsymbol{\theta}_k|\mathcal{M}_k\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathcal{M}_k)$ on each of the K models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathcal{M}_1|\,\mathbf{y}),\ldots,\Pr(\mathcal{M}_K|\mathbf{y}\}$ conditioned on the data but marginalized over the unknowns $\boldsymbol{\theta}_k$ (Jeffreys 1961). Specifically, the posterior probability of $\mathcal{M}_k$ is given by the expression

$$\begin{aligned}\Pr(\mathcal{M}_k|\,\mathbf{y}) &= \frac{\Pr(\mathcal{M}_k)m(\mathbf{y}|\,\mathcal{M}_k)}{\sum_{l=1}^{K}\Pr(\mathcal{M}_l)m(\mathbf{y}|\,\mathcal{M}_l)}\\ &\propto \Pr(\mathcal{M}_k)m(\mathbf{y}|\,\mathcal{M}_k),\quad (k\le K)\end{aligned}$$

where $m(\mathbf{y}|\,\mathcal{M}_k)$ is the marginal likelihood of $\mathcal{M}_k$.

A problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of $\boldsymbol{\theta}_k$. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density. One method for dealing with this difficulty is due to Chib (1995). The starting point is the expression

$$m(\mathbf{y}|\,\mathcal{M}_k) = \frac{f(\mathbf{y}|\,\boldsymbol{\theta}_k,\mathcal{M}_k)p(\boldsymbol{\theta}_k|\,\mathcal{M}_k)}{\pi(\boldsymbol{\theta}_k|\,\mathbf{y},\mathcal{M}_k)}$$

which is an identity in $\boldsymbol{\theta}_k$. From here an estimate of the marginal likelihood on the log-scale is given by

$$\begin{aligned}\log\hat{m}(\mathbf{y}|\,\mathcal{M}_k) = &\log f\left(\mathbf{y}|\,\boldsymbol{\theta}_k^*,\mathcal{M}_k\right)\\ &+\log p\left(\boldsymbol{\theta}_k^*|\,\mathcal{M}_k\right)\\ &-\log\hat{\pi}\left(\boldsymbol{\theta}_k^*|\,\mathbf{y},\mathcal{M}_k\right)\end{aligned}$$

where $\boldsymbol{\theta}_k^*$ denotes an arbitrarily chosen point and $\hat{\pi}\left(\boldsymbol{\theta}_k^*|\,\mathbf{y},\mathcal{M}_k\right)$ is the estimate of the posterior density at that single point. To estimate the posterior ordinate one utilizes the Gibbs output in conjunction with a decomposition of the ordinate into marginal and conditional components. Chib and Jeliazkov (2001) extend this approach for output produced by the M–H algorithm while Basu and Chib (2003) show how the method can be applied in semiparametric models.

In some cases one is interested in a large number of candidate models, each with parameters $\boldsymbol{\theta}_k \in B_k \subseteq R^{d_k}$. In such cases one can get information about the relative support for the contending models from a model space-parameter space MCMC algorithm. In these algorithms, the models are represented by a categorical variable $\mathcal{M}$ which is then sampled along with the parameters of each model. The posterior distribution of $\mathcal{M}$ is computed as the frequency of times each model is visited. Methods for doing this have been proposed by Carlin and Chib (1995) and Green (1995). Both methods are closely related as shown by Dellaportas et al. (2002) and Godsill (2001). Related methods for the problem of variable selection have also been developed starting with George and McCulloch (1993).

See Also

- ▶ Bayesian Econometrics
- ▶ Bayesian Statistics
- ▶ Econometrics
- ▶ Hierarchical Bayes Models
- ▶ Simulation-Based Estimation

Bibliography

Albert, J.H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669–679.

Basu, S., and S. Chib. 2003. Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association* 98: 224–235.

Carlin, B.P., and S. Chib. 1995. Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B* 57: 473–484.

Chan, K.S. 1993. Asymptotic behavior of the Gibbs sampler. *Journal of the American Statistical Association* 88: 320–326.

Chen, M.H., Q.M. Shao, and J.G. Ibrahim. 2000. *Monte Carlo methods in Bayesian computation*. New York: Springer.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90: 1313–1321.

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

Chib, S., and E. Greenberg. 1994. Bayes inference in regression models with ARMA (p,q) errors. *Journal of Econometrics* 64: 183–206.

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

Chib, S., and I. Jeliazkov. 2001. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* 96: 270–281.

Dellaportas, P., J.J. Forster, and I. Ntzoufras. 2002. On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12: 27–36.

Fahrmeir, L., and G. Tutz. 1994. *Multivariate statistical modelling based on generalized linear models*. Berlin: Springer.

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

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions, PAMI* 6: 721–741.

George, E.I., and R.E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88: 881–889.

Gilks, W.K., S. Richardson, and D.J. Spiegelhalter. 1996. *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Godsill, S.J. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* 10: 230–248.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711–732.

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

Jeffreys, H. 1961. *Theory of Probability*. 3rd edn. Oxford: Oxford University Press.

Liu, J.S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *Journal of the American Statistical Association* 89: 958–966.

Liu, J.S. 2001. *Monte Carlo strategies in scientific computing*. New York: Springer.

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

Robert, C.P., and G. Casella. 2004. *Monte Carlo statistical methods*. 2nd edn. New York: Springer.

Roberts, G.O., and S.K. Sahu. 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B* 59: 291–317.

Roberts, G.O., and A.F.M. Smith. 1994. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications* 49: 207–216.

Tanner, M.A., and W.H. Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82: 528–550.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 21: 1701–1762.

Tierney, L., and A. Mira. 1999. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* 18: 2507–2515.

Markov Equilibria in Macroeconomics

Dirk Krueger and Felix Kubler

Abstract

We review the recent literature in macroeconomics that analyses Markov equilibria in dynamic general equilibrium model. After defining the Markov equilibrium concept we first summarize what is known about the existence and uniqueness of such equilibria in models where sequential equilibria can be obtained by solving a suitable social planner

problem. We then discuss the existence problems of Markov equilibria in models where equivalence of equilibrium allocations and solutions to social planner problems cannot be established and review techniques the literature has developed to deal with the existence problem, as well as recent applications of these techniques in macroeconomics.

Keywords

Backward induction; Business cycles; Commitment; Convexity; Dynamic programming; Enforcement frictions; Euler equations; Existence of a Markov equilibrium; Fixed-point theorems; Functional equations; Generalized Markov equilibria; Informational frictions; Lagrange multipliers; Markov chains; Markov equilibria in macroeconomics; Markov processes; Neoclassical growth theory; Optimal taxation; Overlapping generations models; Policy functions; Principle of optimality; Recursive Markov equilibria; Recursive techniques; Reputation; Sequential equilibria; Social planner problem; State spaces; Transition functions

JEL Classifications

D4; D10

We say that a dynamic economy has a Markovian structure (or is Markovian, for short) if the stochastic processes that specify the fundamentals of the economy (such as endowments, preferences and technologies) are Markov processes. Note that deterministic economies are special cases in which the stochastic processes for the fundamentals have degenerate distributions. In many applications attention is restricted to first-order Markov processes in which the probability distributions over fundamentals today are functions exclusively of their values yesterday.

In dynamic economies sequential equilibria are sequences of functions mapping histories of realizations of the stochastic process of the fundamentals into allocations and prices such that all agents in the economy maximize their objectives, given prices, and all markets clear. Under fairly mild conditions (that is, convexity and continuity assumptions on the primitives) such equilibria exist. However, in order to characterize and compute equilibria it is often useful to look for equilibria of a different form.

Recursive Markov equilibria can be characterized by a state space, a policy function and a transition function. The policy function maps the state today into current endogenous choices and prices, and the transition function maps the state today into a probability distribution over states tomorrow (see, for example, the definition in Ljungqvist and Sargent's 2000 textbook). In most of this survey we will use the terms 'Markov equilibria' and 'recursive Markov equilibria' interchangeably; however, below we also consider Markov equilibria which are not recursive and refer to these as 'generalized Markov equilibria'. This characterization leaves open, of course, what the appropriate state variables are that constitute the state space.

Most simply, the state space would consist of the set of possible exogenous shocks governing endowments, preferences and technologies. But, other than in exceptional cases (see, for example, Lucas's 1978 asset pricing application where asset prices are solely functions of the underlying shocks to technology), such a strongly stationary Markov equilibrium does not exist.

In addition to the exogenous shocks, endogenous variables have to be included in the state space to assure existence of a Markov equilibrium. We define as the minimal state space the space of all exogenous shocks and endogenous variables that are payoff-relevant today, in that they affect current production or consumption sets or preferences (see Maskin and Tirole 2001).

We call Markov equilibria with this minimal state space 'simple Markov equilibria'. In the remainder of this article we want to discuss what we know about the existence and uniqueness of such Markov equilibria, both in general and for important specific examples. As it turns out, when equilibria are Pareto efficient, and thus equilibrium allocation can be determined by solving a suitable social planner problem, simple Markov equilibria can be shown to exist under fairly mild conditions. We therefore discuss this case first. On the other

hand, when equilibria are not Pareto efficient – for example, when markets are incomplete or economic agents behave strategically – forward-looking variables often have to be included for a Markov equilibrium to exist; therefore, simple Markov equilibria in the sense defined above do not exist in general. We discuss this case in Section 2.

Markov Equilibria in Economies Where Equilibria Are Pareto Optimal

In this section we discuss the existence and uniqueness of simple Markov equilibria in economies whose sequential market equilibrium allocations can be determined as solutions to a suitable social planner problem. In these economies the problem of proving the existence of a Markov equilibrium reduces to showing that the solution of the social planner can be written as a time-invariant optimal policy function of the minimal set of state variables, as defined above.

This is commonly done by reformulating the optimization problem of the social planner as a functional equation and showing that the optimal Markov policy function generates a sequential allocation which solves the original social planner problem; this is what Bellman (1957) called the principle of optimality. This principle can be established under weak conditions (see Stokey et al. 1989). Equipped with this result, the existence of a Markov equilibrium then follows from the existence of a solution to the functional equation associated with the social planner problem.

If the functional equation can be shown to be a contraction mapping (sufficient conditions for this were provided by Blackwell 1965), then it follows that there exists a unique value function solving the functional equation and an optimal policy correspondence. In addition, the contraction mapping theorem also gives an iterative procedure to find the solution to the functional equation from any starting guess, which is helpful for numerical work.

Under weaker conditions other fixed-point theorems may be employed to argue at least for the existence (if not uniqueness) of a solution to the functional equation, with associated optimal Markov policy correspondence. In order to establish that the policy correspondence is actually a function (and thus the Markov equilibrium is unique), in general strict concavity of the return function needs to be assumed. Stokey et al. (1989) provide a summary of the main results in the general theory of dynamic programming.

This technique of analysing and computing dynamic equilibria in Pareto optimal economies is now widely used in macroeconomics. Its first application can be found in Lucas and Prescott (1971) in their study of optimal investment behaviour under uncertainty. Lucas (1978) used recursive techniques to study asset prices in an endowment economy and showed that the Markov equilibrium has a particularly simple form. Kydland and Prescott (1982) showed how powerful these techniques are for a quantitative study of the business cycle implications of the neoclassical growth model with technology shocks to production. The volume by Cooley (1995) provides a comprehensive overview over this line of research.

M

Generalized Markov Equilibria

In models where the first welfare theorem is not applicable (for example, models with incomplete financial markets or with distorting taxes), in models where there are infinitely many agents (such as overlapping generations models) or in models with strategic interaction the existence of simple Markov equilibria (that is, Markov equilibria with minimal state space) cannot be guaranteed. See Santos 2002; Krebs 2004; Kubler and Schmedders 2002; and Kubler and Polemarchakis 2004, for simple counter-examples. An important exception is Bewley-style models with incomplete markets where simple recursive Markov equilibria exist; see, for example, Krebs 2006. The functional equations characterizing equilibrium have no contraction properties, and more general fixed-point theorems than the contraction mapping theorem, such as Schauder's fixed-point theorem, cannot be applied because it is difficult to guarantee compactness of the space of admissible functions.

Coleman (1991) is an important example where existence can be shown. However, his results rely on monotonicity conditions on the equilibrium dynamics which are not satisfied in general models.

In the applied literature a solution to this problem was suggested early on. For example Kydland and Prescott (1980) analyse a Ramsey dynamic optimal taxation problem. To make the problem recursive they add as a state variable last period's marginal utility.

On the theoretical side Duffie et al. (1994) were the first to rigorously analyse situations where recursive equilibria may fail to exist in general equilibrium models. Kubler and Schmedders (2003) and Miao and Santos (2005) refine their approach and make it applicable for computations. Miao and Santos (2005), also give a clear explanation of how this approach relates to the work by Abreu et al. (1990). We now present their basic idea.

Consider a Markovian economy where a date-event (or node) can be associated with a finite history of shocks, $S^t = (s_0, \ldots, s_t)$. The shocks follow a Markov chain with support $\mathcal{S} = \{1, \ldots, S\}$. Denote by $z(s^t)$ the vector of all endogenous variables at node s^t. Typically this would include the vector of household asset holdings across individuals and the capital stock at the beginning of the period, but also prices and endogenous choices at note s^t, as well as shadow variables such as Lagrange multipliers. A competitive equilibrium is a process of endogenous variables $\{z(s^t)\}$ with $z(s^t) \in \mathcal{Z} \subset \mathbb{R}^M$, which solve the optimization problems of all agents in the economy, and clear markets. The set $\mathcal{Z}$ denotes the set of all possible values of the endogenous variables.

We focus on dynamic economic models where an equilibrium can be characterized by a set of equations relating current-period exogenous and endogenous variables to endogenous and exogenous variables next period. It is straightforward to incorporate inequality constraints into this framework. For expositional purposes we focus on equations. Examples of such equations are the Euler equations of individual households, first order conditions of firms, as well as market-clearing conditions for all markets. We assume that such a set of equations characterizing equilibrium is given and denote it by

$$h(\hat{s}, \hat{z}, z_1, \ldots, z_S) = 0.$$

The arguments $(\hat{s}, \hat{z})$ denote the exogenous state variables and endogenous variables for the current period. Note that the endogenous variables might contain variables which were determined in the previous period, such as the capital stock and individuals' assets. The variables $(z_s)_{s=1}^S$ denote endogenous variables in the subsequent period, in states $s = 1, \ldots, S$, respectively. We refer to $h(\cdot) = 0$ as the set of 'equilibrium equations'.

As explained above, to analyse Markov equilibria one needs to specify an appropriate state space. We assume that the equilibrium set $\mathcal{Z}$ can be written as the product $\mathcal{Y} \times \hat{\mathcal{Z}}$, where $\mathcal{Y}$ denotes the set into which the endogenous state variables fall. In the neoclassical growth model, $\mathcal{Y}$ would consist of the set of possible values of the capital stock; in models with heterogeneous agents one would need to add the set of possible wealth distributions across agents. Unfortunately, as the references cited above show, a recursive Markov equilibrium with this state space may not exist. We therefore require a more general notion of Markov equilibrium for these types of economies.

A *generalized Markov equilibrium* consists of a (non-empty valued) 'policy correspondence', P, that maps the state today into possible endogenous variables today, and a 'transition function', F, that maps the state and endogenous variables today into endogenous variables next period. Formally, the maps

$$P : \mathcal{S} \times \mathcal{Y} \rightrightarrows \hat{\mathcal{Z}} \text{ and } F : \text{graph}(P) \to \mathcal{Z}^S$$

should satisfy that for all shocks and endogenous variables in the current period, $(\hat{s}, \hat{z}) \in \text{graph}(P)$, the transition function prescribes values next period that are consistent with the equilibrium equations, that is,

$$h(\hat{s}, \hat{z}, F(\hat{s}, \hat{z})) = 0,$$

and lie in the policy correspondence, that is,

$$(s, F_s(\hat{s}, \hat{z})) \in \text{graph}(P) \text{ for all } s \in \text{S}.$$

It follows that a generalized Markov equilibrium is recursive, according to our earlier definition, if the associated policy correspondence is single valued. It is simple if the state space is the natural minimal state space.

It is easy to see that Markov equilibria are in fact competitive equilibria in the usual sense. Duffie et al. (1994) show that, under mild assumptions on the primitives of the model, generalized Markov equilibria exist whenever competitive equilibria exist. The basic idea of their approach is very similar to backward induction, using critically a natural monotonicity property of the inverse of the equilibrium equations. (See their original paper, Kubler and Schmedders (2003), or Miao and Santos (2005), for details.)

For practical purposes it is of course crucial that the chosen state space is relatively small and that the Markov equilibrium is recursive. In an asset pricing model with heterogeneous agents, Kubler and Schmedders choose the state space to consist of the beginning-of-period wealth distribution, but can show the existence only of a generalized Markov equilibrium. One cannot rule out the possibility that the equilibrium is not recursive; the same value of the state variables might occur with different values of the endogenous variables. The counter-examples to existence mentioned above show that this is precisely the problem. If for given initial conditions there exist multiple competitive equilibria, the one that realizes is pinned down by lagged variables. Without ruling out multiplicity of equilibria, it does not seem possible to prove the existence of recursive equilibria with the natural state space.

Miao and Santos (2005) enlarge the state space with the shadow values of investment of all agents and prove that with this larger state space a recursive Markov equilibrium exists. The basic insight of their approach is that one needs to add variables to the natural state space that uniquely select one out of several possible endogenous variables.

The main practical problem with the approach originated by Duffie et al. (1994) and refined by Miao and Santos (2005) is that it provides a method to construct all Markov equilibria. There might exist some recursive equilibria for the natural (minimal) state space, but this approach naturally solves for all other recursive Markov equilibria as well. Datta et al. 2005, provide ideas for solving for the one Markov equilibrium with minimal state space.

In many recent applications of recursive methods to macroeconomics the focus of researchers studying non-optimal economies is to find a recursive equilibrium with minimal state space. Notable examples in which even this natural state space is large are Rios-Rull (1996), Heaton and Lucas (1996) and Krusell and Smith (1998). They mark the boundary of economies that currently can be analysed with recursive techniques.

In dynamic endowment economies with either informational frictions or limited enforceability of contracts, constrained-efficient (efficient, subject to the informational or enforcement constraints) consumption allocations usually display a high degree of dependence on past endowment shocks, even though the natural state space contains only the current endowment shock. Therefore, Markov equilibria with minimal state space do not exist. However, using ideas by Spear and Srivastava (1987) and Abreu et al. (1990), the papers by Atkeson and Lucas (1992) and Thomas and Worrall (1988) demonstrate that nevertheless the constrained social planner problem has a convenient recursive structure if one includes promised lifetime utility as a state variable into the recursive problem. This approach or its close alternative, namely, to introduce as an additional state variable Lagrange multipliers on the incentive or enforcement constraints (as in Marcet and Marimon 1998), has seen many applications in macroeconomics, since it facilitates making a large class of dynamic models with informational or enforcement frictions recursive and hence tractable. Miao and Santos (2005) show how such problems with strategic interactions can be incorporated into the framework above.

In optimal policy problems in which the government has no access to a commitment technology, a discussion has emerged about the

desirability of a restriction to Markov policies with minimal state space. Such restrictions rule out reputation if one confines attention to smooth policies. See Phelan and Stacchetti (2001) and Klein and Rios-Rull (2003) for examples of the two opposing views on this issue. However, as Krusell and Smith (2003) argue, if one allows discontinuous policy functions reputation effects can be generated even with Markov policies. (While Krusell and Smith discuss optimal decision rules in a consumption–savings problem with quasi-geometric discounting, their results carry over to optimal policy problems without commitment on the part of the policymaker.)

See Also

- ▶ Computation of General Equilibria
- ▶ Decentralization
- ▶ Euler Equations
- ▶ Existence of General Equilibrium
- ▶ Functional Analysis
- ▶ General Equilibrium
- ▶ General Equilibrium (New Developments)
- ▶ Income Taxation and Optimal Policies
- ▶ Incomplete Markets
- ▶ Markov Processes
- ▶ Optimal Fiscal and Monetary Policy (Without Commitment)
- ▶ Pareto Efficiency
- ▶ Recursive Competitive Equilibrium
- ▶ Recursive Contracts

Bibliography

Abreu, D., D. Pearce, and E. Stacchetti. 1990. Toward a theory of repeated games with discounting. *Econometrica* 58: 1041–1063.

Atkeson, A., and R. Lucas. 1992. On efficient distribution with private information. *Review of Economic Studies* 59: 427–453.

Bellman, R. 1957. *Dynamic programming*. Princeton: Princeton University Press.

Blackwell, D. 1965. Discounted dynamic programming. *Annals of Mathematical Statistics* 36: 226–235.

Coleman, J. 1991. Equilibrium in a production economy with an income tax. *Econometrica* 59: 1091–1104.

Cooley, T. 1995. *Frontiers of business cycle research*. Princeton: Princeton University Press.

Datta, M., L. Mirman, O. Morand, and K. Reffett. 2005. Markovian equilibrium in infinite horizon economies with incomplete markets and public policy. *Journal of Mathematical Economics* 41: 505–544.

Duffie, D., J. Geanakoplos, A. Mas-Colell, and A. McLennan. 1994. Stationary Markov equilibria. *Econometrica* 62: 745–781.

Heaton, H., and D. Lucas. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104: 443–487.

Klein, P., and V. Rios-Rull. 2003. Time consistent optimal fiscal policy. *International Economic Review* 44: 1217–1246.

Krebs, T. 2004. Non-existence of recursive equilibria on compact state spaces when markets are incomplete. *Journal of Economic Theory* 115: 134–150.

Krebs, T. 2006. Recursive equilibrium in endogenous growth models with incomplete markets. *Economic Theory* 29: 505–523.

Krusell, P., and A. Smith. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy* 106: 867–896.

Krusell, P., and A. Smith. 2003. Consumption–savings decisions with quasi–geometric discounting. *Econometrica* 71: 365–376.

Kubler, F., and H. Polemarchakis. 2004. Stationary Markov equilibria for overlapping generations. *Economic Theory* 24: 623–643.

Kubler, F., and K. Schmedders. 2002. Recursive equilibria in economies with incomplete markets. *Macroeconomic Dynamics* 6: 284–306.

Kubler, F., and K. Schmedders. 2003. Stationary equilibria in asset-pricing models with incomplete markets and collateral. *Econometrica* 71: 1767–1795.

Kydland, F., and E. Prescott. 1980. Dynamic optimal taxation, rational expectations and optimal control. *Journal of Economic Dynamics and Control* 2: 79–91.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1371.

Ljungqvist, L., and T. Sargent. 2000. *Recursive macroeconomic theory*. Cambridge, MA: MIT Press.

Lucas, R. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1426–1445.

Lucas, R., and E. Prescott. 1971. Investment under uncertainty. *Econometrica* 39: 659–681.

Marcet, A. and Marimon, R. 1998. Recursive contracts. Working paper, University Pompeu Fabra, Barcelona.

Maskin, E., and J. Tirole. 2001. Markov perfect equilibrium. *Journal of Economic Theory* 100: 191–219.

Miao, J., and Santos, M. 2005. Existence and computation of Markov equilibria for dynamic non–optimal economies. Working paper, Department of Economics, Boston University.

Phelan, C., and E. Stacchetti. 2001. Sequential equilibria in a Ramsey tax model. *Econometrica* 69: 1491–1518.

Rios-Rull, V. 1996. Life cycle economies with aggregate fluctuations. *Review of Economic Studies* 63: 465–490.

Santos, M. 2002. On non-existence of Markov equilibria for competitive-market economies. *Journal of Economic Theory* 105: 73–98.
Spear, S., and S. Srivastava. 1987. On repeated moral hazard with discounting. *Review of Economic Studies* 54: 599–617.
Stokey, N., R. Lucas, and E. Prescott. 1989. *Recursive methods in economic dynamics*. Cambridge, MA: Harvard University Press.
Thomas, J., and T. Worrall. 1988. Self-enforcing wage contracts. *Review of Economic Studies* 55: 541–554.

Markov Processes

Daniel W. Stroock

Abstract

In this article the theory of Markov processes is described as an evolution on the space of probability measures. Following a brief historical account of its origins in physics, a mathematical formulation of the theory is given. Emphasis has been placed on the ergodic properties of Markov processes, and their presence is checked in a simple example.

Keywords

Boltzmann, L.; Brownian motion; Ergodic theory; Gibbs, G.; Markov processes; Markov property; Newton's equation; Probability; Statistical mechanics; Wiener process

JEL Classifications

C6

Unless one is clairvoyant, the only temporally evolving processes which are tractable are those whose future behaviour can be predicted on the basis of data which is available at the time when the prediction is being made. Of course, in general, the behaviour of even such an evolution will be impossible to predict. For example, if, in order to make a prediction, one has to know the detailed history of everything that has happened during the entire history of the entire universe, one's chance of making a prediction may be a practical, if not a theoretical, impossibility. For this reason, one tries to study evolutions mathematically with models in which most of the distant past can be ignored when one makes predictions about the future. In fact, many mathematical models of evolutions have the property that, for the purpose of predicting the future, the past becomes irrelevant as soon as one knows the present, in which case the evolution is said to be a 'Markov process', the topic at hand, after Andrei Andreyevich Markov (1856–1922).

The components of a Markov process are its *state space* S and its *transition rule T*. Mathematically, S is just some non-empty set, which in applications will encode all the possible states in which the evolving system can find itself, and $T : S \to S$ is a function from S into itself which gives the *transition rule*. More precisely, if now the system is in state x, it will be next in state $T(x)$, from which it will go to $T^2(x) = T(T(x))$, and so on. (Here we are thinking of time being discrete. Thus, 'next' means after one unit of time has passed.)

To give a sense of the sort of reasoning required to construct a Markov process, consider a (classical) physical particle whose motion is governed by Newton's equation $\Rightarrow F = m \Rightarrow a$ ('force equals mass times acceleration'). At least in theory, Newton's equation says that, on the assumption that one knows the mass of the particle and the force field $\Rightarrow F$ which acts on it, one can predict where the particle will be in the future as soon as one knows what its position and velocity are now. On the other hand, knowing only its present position is not sufficient by itself. Thus, even though one may care about nothing but its position, in order to produce a Markov process for a particle evolving according to Newton's equation it is necessary to adopt the attitude that the *state* of the particle consists of its position *and* velocity, not just its position alone. Of course, in that velocity is the derivative of position, the two are so inextricably intertwined that one might be tempted to concentrate on position on the grounds that one will be able to compute the velocity whenever necessary. However, this tack destroys the Markov property, namely, there is no

way of computing the velocity of a particle 'now' if all one knows is its position 'now'. For this reason, physicists consider the state of a particle to be a composite of its position and velocity, and the resulting state space $\mathbb{R}^6 = \mathbb{R}^3 \times \mathbb{R}^3$ (three coordinates for position and three for velocity) they call the *phase space* of the particle.

The same point may be clearer in the following example. Suppose that one has an evolution on a state space S which proceeds according to the rule that, if the present state is x_n and the preceding state was x_{n-1}, then the next state will be $x_{n+1} = T(x_{n-1}, x_n)$. This is *not* a Markov process. Nonetheless, it can be 'Markovized'. Indeed, replace the original state space by $\hat{\mathbb{S}} = \mathbb{S} \times \mathbb{S}$, the set of ordered pairs (x, y) with x and y from S, and define $\hat{T}((x,y)) = (y, T(x,y))$. It is then an easy matter to check that, if the original system was in state x_{-1} at time -1 and state x_0 at time 0, then its state at time $n \geq 1$ will be x_n, the second component of the pair $(x_{n-1}, x_n) = \hat{T}^n((x_{-1}, x_0))$.

The moral to be drawn from these examples is that *the presence or absence of the Markov property is in the eye of the beholder*. That is, a change of venue (the state space) can make the Markov property appear in circumstances where it was not originally apparent. In fact, by making the state space sufficiently large, any evolution can be forced to be Markov. On the other hand, the more complicated the state space, the less useful is the Markov property. Thus, in practice, what one seeks is the 'simplest' state space on which one's evolution possesses the Markov property.

Stochastic Markov Processes

Roughly speaking, Markov processes fall into one of two categories. Those in the first category are 'deterministic' in the sense that their state space is sufficiently detailed that the individual states give complete and unambiguous information. Both the examples given above are deterministic. The mathematical analysis of deterministic Markov processes has a proud history going back to Newton which includes major contributions by such luminaries as P. Chebyshev, A. Markov, A. Lyapounov, H. Poincaré, and J. Moser. The second category of Markov processes, and the one on which the rest of this article will concentrate, are 'probabilistic' or 'stochastic' Markov processes. To understand where and why these processes arise, consider the problem of describing the state of all the gas molecules in a room. Each litre of gas contains approximately Avogrado's number, $6.02214199 \times 10^{23}$, of molecules. Thus, even a small room will contain something on the order of 10^{26} molecules. Moreover, because, by Newton's laws of motion, the state of each individual molecule will lie in its individual phase space, the state of the entire system of molecules will have to specify the positions and velocities of all 10^{26} molecules. Stated mathematically, the state space of the system will be $\mathbb{R}^{6\times 10^{26}}$, on which any sort serious analysis is too daunting to contemplate.

When one is confronted with a problem which is intractable as presented, the time-honoured procedure of choice is to reformulate the problem in a way which makes it more tractable. In the case just described, the reformulation was made by G.W. Gibbs (1902) and L. Boltzmann (1896, 1898), the fathers of statistical mechanics. They abandoned any hope of saying exactly where all the molecules will be and reconciled themselves to settling for a description of the statistics of the molecules. That is, instead of asking exactly where all the molecules would be, they asked what would be the probability of finding a molecule in various regions of phase space. From this point of view, the state of the system will not be an element of $\mathbb{R}^{6\times 10^{26}}$ but of $\mathbf{M_1}(\mathbb{R}^6)$, the space probability distributions on the individual phase space $\mathbb{R}^6$. Of course, Gibbs and Boltzmann's reformulation only changes the problem, it does not solve it. Indeed, although Newton's equation determines how the system of molecules evolves and therefore how their distribution will evolve, the use of Newton's equation would remove the advantage which Boltzmann and Gibbs hoped to gain from their reformulation. Thus, they had to come up with an alternative way of describing the transition rule which governs the evolution of the distribution of the system as a Markov process on $\mathbf{M_1}(\mathbb{R}^6)$. The description proposed by Boltzmann is given by the famous Boltzmann equation.

Unfortunately, Boltzmann's equation is itself so complicated that it is only recently that substantial progress has been made toward understanding it in any generality. On the other hand, Gibbs and Boltzmann's idea of studying Markov processes on the space of probability distributions is seminal and has proved to be both ubiquitous and powerful.

The abstract setting for a stochastic Markov process starts with a non-empty set S, the deterministic state space, and the associated space $\mathbf{M}_1(\mathrm{S})$ of probability distributions on S. The easiest and most commonly studied stochastic Markov processes are those for which the transition rule $T : \mathbf{M}_1(\mathrm{S}) \to \mathbf{M}_1(\mathrm{S})$ is a linear (more correctly, an affine) function. To be definite, suppose S is a finite set. Then $\mathbf{M}_1(\mathrm{S})$ is the set of all functions μ on S which assign each $x \in \mathrm{S}$ a number $\mu(\{x\}) \in [0, 1]$ (the probability of $\{x\}$ under μ) in such a way that $\sum_{x \in \mathrm{S}} \mu(\{x\}) = 1$. (The use of $\mu(\{x\})$ instead of $\mu(x)$ here is a little pedantic. However, one must remember that probabilities are assigned to *events* – that is, subsets of the sample space – and that $\{x\}$ is the event that 'x occurred'.) Clearly, if μ and ν are in $\mathbf{M}_1(\mathrm{S})$ and $\theta \in [0, 1]$, then the convex combination $\theta\mu + (1 - \theta)\nu$ is again an element of $\mathbf{M}_1(\mathrm{S})$. Sets with this property are said to be 'affine' (as distinguished from 'linear', which refers to sets which are closed under all linear, not just convex, combinations), and a function on an affine set is said to be affine if it commutes with convex combinations. Thus, for $\mathbf{M}_1(\mathrm{S})$, the transition rule T is affine if $T(\theta\mu + (1 - \theta)\nu) = \theta T(\mu) + (1 - \theta)T(\nu)$. Because S is finite, one can dissect such transition rules in the following way. First, for each $x \in \mathrm{S}$, let δ_x denote the element of $\mathbf{M}_1(\mathrm{S})$ which assigns 1 to $\{x\}$ (and therefore 0 to $\mathrm{S}/\{x\}$). Next, set $\mathbf{P}(\mathrm{x}, \cdot) = T(\delta_{\mathrm{x}})$. That is, $\mathbf{P}(\mathrm{x}, \cdot)$ is the element of $\mathbf{M}_1(\mathrm{S})$ to which T takes $T(\delta_x)$, and so $\mathbf{P}(x, \{y\}) = [T(\delta_x)](\{y\})$. Because, for any $\mu \in \mathbf{M}_1(\mathrm{S})$ which is not equal to δ_x, $\mu = \mu(\{x\})\delta_x) + (1 - \mu(\{x\}))\mu^x$, where $\mu^x \in \mathbf{M}_1(\mathrm{S})$ is determined so that $\mu^x(\{y\})$ equals $(1 - \mu(\{x\}))^{-1}\mu(\{y\})$ or 0 depending on whether $y \neq x$ or $y = x$, the affine property of T means that $T(\mu) = \mu(\{x\})\mathbf{P}(x, \cdot) + (1 - \mu(\{x\}))T(\mu^x)$. Hence, after peeling off one x at a time, one concludes that

$$T(\mu) \sum_{x \in \mathbb{S}} \mu(\{x\})\mathbf{P}(\mathrm{x}, \cdot) \tag{1}$$

when T is affine.

Probabilistic Interpretation

The representation of T given by (1) admits an intuitively pleasing probabilistic interpretation: namely, $\mathbf{P}(\mathrm{x}, \{\mathrm{y}\})$ can be thought of as the probability that the system will next be in the state y given that is now in state x. With this interpretation in mind, probabilists call $x \in \mathrm{S} \mapsto \mathbf{P}(\mathrm{x}, \cdot) \in \mathbf{M}_1(\mathrm{S})$ a *transition probability* on the state space S. The terminology here is confusing. From the point of view adopted earlier, one might, and should, have thought that $\mathbf{M}_1(\mathrm{S})$ is the state space. However, the probabilistic interpretation is most easily appreciated if one thinks of S as the state space and $x \in \mathrm{S} \to (\mathrm{x}, \cdot) \in \mathbf{M}_1(\mathrm{S})$ as a random transition rule. To complete this picture, probabilists introduce random variables to represent the random points in S visited. More precisely, again assume that S is finite, and suppose that $\mu \in \mathbf{M}_1(\mathrm{S})$ describes the initial distribution of the process under consideration. Then probabilists construct a sequence $\{X_n : n \geq 0\}$ of random variables, called a *Markov chain*, in such a way that, for any $n \geq 0$,

$$\begin{aligned}&\mathbf{P}(X_0 = x_0, \ldots, X_n = x_n)\\ &\quad = \mu(\{x_0\})\mathbf{P}(\mathrm{x}_0, \{\mathrm{x}_1\}) \cdots \mathbf{P}(\mathrm{x}_{\mathrm{n}-1}, \{\mathrm{x}_\mathrm{n}\}).\end{aligned}$$

In words, this says that the right-hand side above is the probability that the chain with initial distribution μ starts at x_0 and then goes on to visit, successively, the points x_1 through x_n.

To see that the probabilistic interpretation is completely consistent in the deterministic case, observe that a deterministic Markov process can be formulated as a stochastic Markov process. That is, if T is the transition rule for the deterministic process, take $\mathbf{P}(\mathrm{x}, \cdot) = \delta_{\mathrm{T(x)}}$, and check that, with probability 1, the Markov chain with transition probability $\mathbf{P}(\mathrm{x}, \cdot)$ follows the same path as the deterministic one with transition rule T. Equivalently, with probability 1, $X_n = T^n(X_0)$ for all $n \geq 1$.

Ergodic Theory of Markov Chains

Continue in the setting of the preceding section. One of the phenomena predicted by Gibbs in connection with his and Boltzmann's study of gases was that, no matter what the initial distribution of the gas, after a long time the gas should equilibrate in the sense that it will achieve a *stationary distribution* (that is, a distribution that does not change with time) which does not depend on how it was distributed initially. One's experience with the behaviour of gases makes this prediction entirely plausible: place an opened bottle of perfume in the corner of a room; wait an hour, and confirm that the perfume will have become more or less equidistributed throughout the room. Be that as it may, the prediction, which goes by the name of Gibbs's 'ergodic hypothesis', has been mathematically verified in only one physically realistic model. Nonetheless, as will be explained next, ergodicity is relatively easy to verify for most stochastic Markov processes on a finite state space.

To develop some intuition for what ergodicity means and why it might hold for a stochastic Markov process on a finite state space S, it is best to first know how to recognize when a $\mu \in \mathbf{M}_1(\mathrm{S})$ is stationary. But, if μ is stationary, then it is left unchanged as the system evolves, and, in terms of the transition probability, this means that

$$\mu(\{y\}) = \sum_{x \in \mathbb{S}} \mu(\{x\})\mathbf{P}(\mathrm{x}, \{\mathrm{y}\}) \text{ for all } \mathrm{y} \in \mathbb{S}. \tag{2}$$

Now suppose that $\mathrm{S} = \{1, 2\}$, and consider the problem of finding a solution to (2). That is, we want to find $\mu \in \mathbf{M}_1(\{1, 2\})$ so that

$$\begin{aligned} \mu(\{1\}) &= \mu(\{1\})\mathbf{P}(1, \{1\}) \\ &\quad + \mu\big(\{2\}\mathbf{P}(2, \{1\})\mu(\{2\}) \\ &= \mu(\{1\})\mathbf{P}(1, \{2\}) + \mu\big(\{2\}\mathbf{P}(2, \{2\}). \end{aligned} \tag{3}$$

At first sight, there appear to be too many conditions on μ: not only must it satisfy the two equations in (3), it also has to satisfy $\mu(\{1\}) + \mu(\{2\}) = 1$ as well as being non-negative. Even if one ignores the non-negativity, one suspects that three linear equations are just too many for a pair of numbers to satisfy. On the other hand, after a little manipulation, one sees (remember that $\mathbf{P}(1, \cdot)$ and $\mathbf{P}(2, \cdot)$ are probability distributions) that both the equations in (3) are equivalent to $\mu(\{1\})\mathbf{P}(1, \{2\}) = \mu(\{2\})\mathbf{P}(2, \{1\})$. Hence the two equations in (3) are equivalent, and so there are really only two equations to be satisfied: $\mu(\{1\})\mathbf{P}(1, \{2\}) = \mu(\{2\})\mathbf{P}(2, \{1\})$ and $\mu(\{1\}) + \mu(\{2\}) = 1$. There are two cases to be considered. The first case is when the chain never moves, or, equivalently, $\mathbf{P}(1, \{2\}) = 0 = \mathbf{P}(2, \{1\})$. In this case there are two solutions, namely, δ_1 and δ_2, which is exactly what one should expect for a chain which never moves. In the second case, the one corresponding to a chain which can move, either $\mathbf{P}(1, \{2\}) > 0$ or $\mathbf{P}(2, \{1\}) > 0$. In both these cases, one can easily check that the one and only solution to (3) is given by

$$\begin{aligned} \mu(\{1\}) &= \frac{\mathbf{P}(2, \{1\})}{\mathbf{P}(1, \{1\}) + \mathbf{P}(2, \{1\})} \text{ and } \mu(\{2\}) \\ &= \frac{\mathbf{P}(1, \{2\})}{\mathbf{P}(1, \{2\}) + \mathbf{P}(2, \{1\})}. \end{aligned}$$

Continuing in the setting of the preceding, we want to examine when Gibbs's ergodic hypothesis holds. Obviously, at the very least, ergodicity requires that there be only one stationary μ, otherwise we could start the chain with one of them as initial distribution, in which case it would never get to the other. Thus, we need to assume that $\mathbf{P}(1, \{2\}) + \mathbf{P}(2, \{1\}) > 0$, and, to simplify matters, we will assume more, namely, that $m \equiv m_1 + m_2 > 0$, where $m_1 = \min\{P(1, \{1\}), P(2, \{1\})\}$ and $m_2 = \min\{P(1, \{2\}), P(2, \{2\})\}$, and, under this assumption we (following Doeblin) will show that, for any $\nu \in M_1(\{1, 2\})$

$$\|\nu\mathbf{P} - \mu\| \leq (1 - m)\|\nu - \mu\|, \tag{4}$$

where $\nu\mathbf{P} \in \mathbf{M}_1(\{1, 2\})$ is determined by

$$\nu\mathbf{P}(\{\mathrm{y}\}) \sum_{\mathrm{x}=1}^{2} \nu(\{\mathrm{x}\})\mathbf{P}(\mathrm{x}, \{\mathrm{y}\})$$

and, for any pair $\nu_1, \nu_2 \in \mathbf{M}_1(\{1,2\})$, $\|\nu_2 - \nu_1\| \equiv \sum_{x=1}^{2} |\nu_2(\{x\}) - \nu_1(\{x\})|$. To prove (4), first observe that, because μ is stationary, $\mu = \mu\mathbf{P}$, and therefore, since $\sum_{x=1}^{2} (\nu(\{x\}) - \mu(\{x\})) = 1 - 1 = 0$,

$$\nu P(\{\mathrm{y}\}) - \mu(\{\mathrm{y}\}) = \sum_{\mathrm{x}=1}^{2} (\nu(\{\mathrm{x}\}) - \mu(\{\mathrm{x}\}))$$
$$\mathbf{P}(\mathrm{x}, \{\mathrm{y}\}) = \sum_{\mathrm{x}=1}^{2} (\nu(\{\mathrm{x}\}) - \mu(\{\mathrm{x}\}))$$
$$\left(\mathbf{P}(\mathrm{x}, \{\mathrm{y}\}) - m_{\mathrm{y}}\right).$$

Next, take the absolute value of both sides, remember that the absolute value of a sum of numbers is dominated by the sum of their absolute values, and arrive at

$$\begin{aligned}
&\|\nu P - \mu\| \\
&\leq \sum_{\mathrm{y}=1}^{2} \left(\sum_{\mathrm{x}=1}^{2} |\nu(\{\mathrm{x}\}) - \mu(\{\mathrm{x}\})| \mathbf{P}\left(\mathrm{x}, \{\mathrm{y}\} - m_{\mathrm{y}}\right) \right) \\
&= \sum_{\mathrm{x}=1}^{2} \left(\sum_{\mathrm{y}=1}^{2} |\nu(\{\mathrm{x}\}) - \mu(\{\mathrm{x}\})| \mathbf{P}\left(\mathrm{x}, \{\mathrm{y}\} - m_{\mathrm{y}}\right) \right) \\
&= (1 - m)\|\nu - \mu\|.
\end{aligned}$$

Given (4), it becomes an easy matter to check ergodicity. Indeed, νP is the distribution of the chain at time 1 when it is started with initial distribution ν. Similarly, its distribution at time 2 will be $\nu\mathbf{P}^2 = (\nu P)\mathbf{P}$, and so: $\|\nu\mathbf{P}^2 - \mu\| \leq (1 - m)\|\nu P - \mu\| \leq (1 - m)^2\|\nu - \mu\|$. Proceeding by induction, one sees that distribution $\nu\mathbf{P}^{\mathbf{n}} = (\nu\mathbf{P}^{\mathbf{n}-1})\mathbf{P}$ at time n will satisfy $\|\nu\mathbf{P}^{\mathbf{n}} - \mu\| \leq (1 - m)^{\mathbf{n}}\|\nu - \mu\|$. Hence, because $m > 0$, this implies that $\|\nu\mathbf{P}^{\mathbf{n}} - \mu\|$ tends to 0 exponentially fast, which means that the chain possesses an extremely strong form of ergodicity.

Other Directions

In this article we have discussed only the most elementary examples of Markov processes. In particular, in order to avoid technical difficulties, all our considerations have been about processes for which the time parameter is discrete. As soon as one moves into the realm of processes with a continuous time parameter, the theory becomes much more technically involved. However, the price which one has to pay in technicalities is amply rewarded by the richness of the continuous time theory. To wit, Brownian motion (also known as the Wiener process) is a continuous parameter Markov process which makes an appearance in a surprising, and ever growing, number of places: harmonic analysis in pure mathematics, filtering and separation of signal from noise in electrical engineering, the kinetic theory of gases in physics, price fluctuations on the stock market in economics, and so on. Thus, for the sake of the curious, the bibliography below gives a very brief and enormously inadequate list of places where one can learn more about Markov processes.

Bibliography

Elementary Texts

Karlin, S., and H. Taylor. 1975. *A first course in stochastic processes*. 2nd ed. New York: Academic Press.

Norris, J. 1997. *Markov chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Stroock, D. 2005. *An introduction to Markov processes*. Graduate Text Series No. 230. Heidelberg: Springer-Verlag.

Advanced Texts

Dynkin, E. 1965. *Markov processes*, vols. 1 and 2. Grundlehren Nos. 121 and 122. Heidelberg: Springer-Verlag.

Ethier, S., and T. Kurtz. 1986. *Markov processes: Characterization and convergence*. New York: Wiley.

Revuz, D. 1984. *Markov chains*. North-Holland Mathematical Library, vol. 11. Amsterdam and New York: North-Holland.

Stroock, D. 2003. *Markov processes from K. Itô's perspective*. Annals of Mathematical Studies No. 155. Princeton: Princeton University Press.

Physics Texts

Boltzmann, L. 1896, 1898. *Lectures on gas theory*,vols. 2, Trans. S. Brush. New York: Dover Publications, 1995.

Gibbs, J. 1902. *Elementary principles in statistical mechanics*. New York: Scribner.

Markowitz, Harry Max (Born 1927)

Donald D. Hester

Abstract

Harry M. Markowitz shared the 1990 Nobel Memorial Prize in Economics with Merton Miller and William Sharpe for their contributions to financial economics. He is principally known for his Cowles Foundation monograph, *Portfolio Selection: Efficient Diversification of Investments*, in which he developed and made accessible to general readers the concept of an efficient portfolio, that is, a collection of assets that has a maximum rate of return for an arbitrary rate of return variance. The monograph provided a rigorous justification for portfolio diversification. He has also developed important applied mathematical tools for working with sparse matrices and performing simulations.

Keywords

American Finance Association; Capital asset pricing model (CAPM); Cholesky factorizations; Customary wealth; Expected utility hypothesis; Linear programming; Markowitz, H.; Portfolio selection; Quadratic utility function; Risk

JEL Classifications

B31

Harry M. Markowitz is a Nobel laureate who shared a 1990 prize with Merton Miller and William Sharpe for their contributions to financial economics. A native of Chicago, he received undergraduate and graduate degrees from the University of Chicago, culminating in a Ph.D. in 1954. His article on portfolio selection (1952a), drawn from his dissertation, was a path-breaking contribution that would be fully developed in his 1959 Cowles Foundation monograph, *Portfolio Selection: Efficient Diversification of Investments*. The monograph provided a strong case for receiving the Nobel Memorial Prize.

Markowitz is a gifted applied mathematical economist who responds creatively to observed behaviour and has a strong interest in providing tools that facilitate applications of economics. As a graduate student he published a second influential article (1952b), which extended and qualified an important contribution by Friedman and Savage (1948) that proposed an explanation for why individuals both insure and gamble. Specifically, he transformed their argument to describe bets that involved deviations from an individual's 'customary wealth', which is wealth exclusive of recent windfall gains or losses, and imposed a third inflection point, which was needed to satisfy the expected utility hypothesis requirement that a utility function be bounded from below. By describing how the Friedman and Savage model could not account for some commonly observed behaviour, this article afforded a clear insight into the way Markowitz analysed decisions about risk. It takes only one simple division to transform deviations from an individual's customary wealth to rates of return on customary wealth.

Markowitz's article on portfolio selection lucidly explained why focusing on the expected rate of return (hereafter, 'return' means 'rate of return') was inadequate to account for widely observed portfolio diversification. By simultaneously considering expected return and the variance of return (E and V), he developed a set of efficient EV portfolios that would have a maximum return for an arbitrary variance of return. Further, almost all of these efficient portfolios would have more than one asset and thus be diversified. Using elegant geometric arguments, the article explained how in a problem involving N securities the set of efficient portfolios could be represented by a set of connected line segments. This insight underlies the algorithm for computing efficient portfolios that is presented in his monograph.

In that article and in his monograph, Markowitz was careful to emphasize that he was developing a method for using an investor's beliefs (or perhaps those of security analysts) about expected return and variance so that he or she could use them in an optimal way. In neither did he explain how expectations should be formed. Similarly, he was agnostic about whether the probabilities investors used in

forming expectations were objective or subjective. Finally, he did not assume that returns were normally distributed or that an investor had a quadratic utility function, although one of these conditions is formally necessary to describe portfolio choice in terms of expected return and variance of return. The complications raised in the preceding three sentences are briefly considered in the final section of the monograph and would absorb many journal pages in the coming years. Levy and Markowitz (1979) addressed the limitations of restricting attention to expected return and variance and argued that by focusing on these two measures investors were not likely often to be misled.

The monograph was an expositional tour de force and consequently had an enormous impact on the theory and practice of finance. Its first chapters were quite intuitive and made no technical demands on the reader. The third and fourth chapters contain elementary discussions of the concepts of expected return and variance, the fifth and sixth generalize the discussion to cover large numbers of securities and aggregation over time, and the seventh provides a clear geometric interpretation of efficient portfolios. The eighth chapter presents the critical line method for isolating efficient portfolios and solving the underlying quadratic programming problem. The ninth chapter restates the argument using a semivariance. The remaining four chapters describe rational portfolio behaviour and discuss how the expected utility hypothesis can be applied to the portfolio selection problem. They include the topics of portfolio choice over time and when objective and subjective probabilities differ.

Technical derivation of the critical line method is reported in Appendix A in the monograph, which generalizes its original exposition in Markowitz (1956). The method works because the set of efficient portfolios is convex, in part because there are assumed to be upper and lower bounds on the holdings of any asset. In the monograph no short sales are allowed, although this restriction can be relaxed. If we ignore some minor technical issues involving singularities that cannot be dealt with here, the method can be described intuitively. It is initiated by finding the security with the highest expected return. A portfolio fully invested in this security is an element in the set of efficient portfolios. Then, find the security or linear combination of securities that can be substituted for that highest-yielding security in a manner which respects the balance sheet identity that the sum of asset shares equals unity and provides the minimum reduction in return per unit decrease in variance. This substitution is continued until one or more of the securities reach zero (or a lower bound) or until a security not in this combination can be beneficially introduced, at which point another linear combination is chosen. The algorithm stops when no substitution is possible that further lowers the variance of a portfolio.

The monograph and a contemporaneous paper by Tobin (1958) underlay the development of the capital asset pricing model (CAPM) by Sharpe (1964), which argued that an asset's return was determined by its correlation with the return of the market portfolio. This model greatly increased interest in EV models among practitioners and the academic community. In his presidential address to the American Finance Association, however, Markowitz (1983) expressed some reservations about the CAPM because it failed to take into account limits on borrowing.

Apart from his work on modelling portfolio decisions, Markowitz made significant contributions to management science. In Markowitz (1957a), he developed sparse matrix techniques for simplifying the solution of linear programming problems, which continue to be used in present-day algorithms that employ Cholesky factorizations. In Markowitz and Manne (1957b) an important set of applications, discrete programming problems, were analysed. In Manne and Markowitz (1963b), applications of 'process analysis' are reported in which Markowitz was a co-author on several papers that studied metal-working industries. Process analysis examines production capabilities in an industry. Also, he made many contributions that led to improvements in simulations, including the construction of a programming language, SIMSCRIPT (see Markowitz, Hausner and Karr 1963a, and Dimsdale and Markowitz 1999).

Markowitz spent much of his career outside academia. From 1952 through 1963 he was on

the staff of the RAND Corporation and from 1974 through 1983 he was at IBM's T. J. Watson Research Center. His monograph was largely written when he was a visitor at Yale University in 1955–6, on leave from RAND. He joined the faculty of Baruch College of the City University of New York as a distinguished professor of finance and economics in 1982, and in 2004 was a research professor at the University of California at San Diego. In 1989 he was awarded the prestigious Von Neumann Prize in Operations Research by the Operations Research Society of America and The Institute of Management Science for his work on 'portfolio selection, mathematical programming, and simulation'.

See Also

- ▶ Computational Methods in Econometrics
- ▶ Efficiency Bounds
- ▶ Expected Utility Hypothesis
- ▶ Foreign Direct Investment
- ▶ Risk
- ▶ Risk-Coping Strategies
- ▶ Sharpe, William F. (Born 1934)
- ▶ Tobin, James (1918–2002)
- ▶ Uncertainty

Selected Works

1952a. Portfolio selection. *Journal of Finance* 7, 77–91.

1952b. The utility of wealth. *Journal of Political Economy* 60(2), 151–158.

1956. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* 3, 111–133.

1957a. The elimination form of the inverse and its application to linear programming. *Management Science* 3, 255–269.

1957b. (With A. Manne.) On the solution of discrete programming problems. *Econometrica* 25, 84–110.

1959. *Portfolio selection: Efficient diversification of investments*. Cowles Foundation Monograph 16. New York: John Wiley and Sons.

1963a. (With B. Hausner and H. Karr.) *SIMSCRIPT: A simulation programming language*. Englewood Cliffs: Prentice Hall.

1963b. (Co-edited with A. Manne.) *Studies in process analysis: Economy-wide production capabilities*. Cowles Foundation Monograph 18. New York: John Wiley and Sons.

1979. (With H. Levy.) Approximating expected utility by a function of mean and variance. *American Economic Review* 69, 308–317.

1983. Negative or not nonnegative: A question about CAPMs. *Journal of Finance* 38, 283–295.

1999. (With B. Dimsdale.) A description of the SIMSCRIPT language. *IBM Systems Journal* 38 (2/3), 151–161.

Bibliography

Friedman, M., and L. Savage. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56: 279–304.

Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Marriage and Divorce

Yoram Weiss

Abstract

We document the increase in marital turnover and survey economic models of the marriage market. Couples match based on attributes but sorting is constrained by costs of search. Divorce is caused by new information on match quality, and remarriage requires further search. Although most men and women marry, they are single more often than before and more children live in one-parent household. The impact on children depends on child-support transfers. Such transfers may rise with the aggregate divorce (remarriage) rates.

Keywords

Altruism; Assortative matching; Child care; Collective models; Commitment; Comparative advantage; Complementarity; Division of labour; Household production; Increasing returns; Leisure; Marriage and divorce; Marriage market; Matching model; Multiple equilibria; Poisson process; Reservation utility; Search models; Sharing rules; Stable sharing rule; Time use; Transferable utility; Unitary models of the household

JEL Classifications

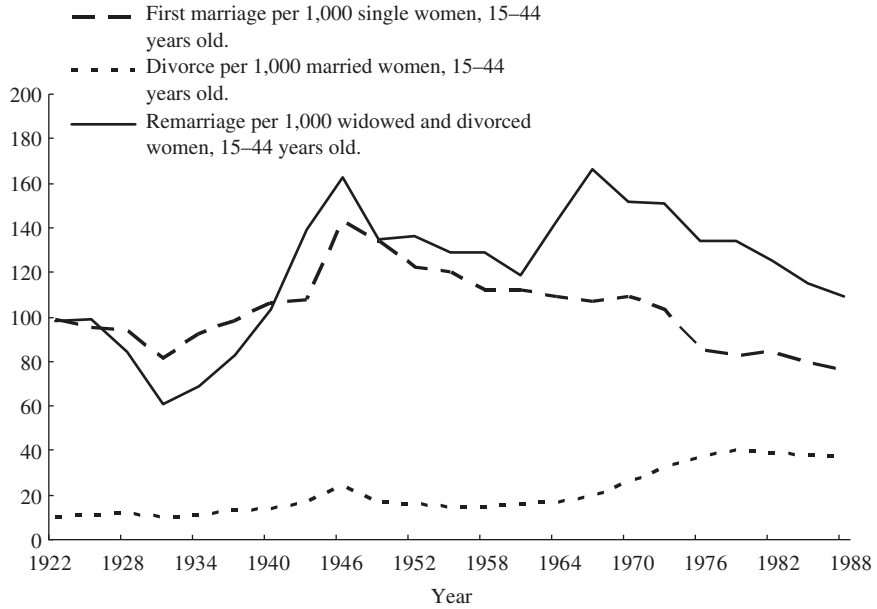
J12

This article summarizes the economic analysis of marriage markets. The first section provides a description of stylized facts that motivate the interest of economists in this problem. It is shown that marital status is closely tied with 'economic' variables such as work and wages. We illustrate these facts using mainly US data but the patterns are similar in all developed countries. The second section demonstrates how the tools of economists bear on 'non-economic' subjects such as marriage, fertility and divorce, often analysed by researchers from other fields. The final section highlights some connections between the theory and empirical evidence.

Basic Facts

Marriage and Divorce

The 20th century was characterized by substantial changes in family structure (Fig. 1). More men and women are now divorced and unmarried or have alternative arrangements, such as cohabitation. Interestingly, the rise in divorce rates is associated with an increase in remarriage rates (relative to first marriage rates), reflecting higher turnover. Most people had a first marriage, and most divorces end in remarriage. Moreover, the remarriage rate is greater than the first marriage rate and far exceeds the divorce rate, suggesting that, despite the larger turnover, marriage is still a 'natural' state (Table 1). Women enter the first marriage faster than men. However, following divorce, men remarry at higher rates than women, especially at



Marriage and Divorce, Fig. 1 Annual numbers (per 1,000) of first marriage, divorce, and remarriage: United States, 1921–1989 (three-year averages). *Source*: National Center of Health Statistics

Marriage and Divorce, Table 1 Marital histories of men and women, United States, 1996

	Ever married by 1966 (%)		Divorced from first marriage by 1966 (%)		Remarried after first divorce by 1966 (%)	
Age in 1966	Men	Women	Men	Women	Men	Women
25	31.8	50.0	4.6	12.2	55.5	44.0
30	65.4	71.1	16.7	17.2	35.6	49.7
35	77.4	84.1	26.9	26.4	60.7	65.1
40	80.9	85.2	34.0	36.5	66.4	67.6
45	87.3	89.8	41.1	41.6	71.6	68.1
50	93.2	91.3	39.8	42.4	78.3	68.9
55	94.5	95.3	38.2	38.0	79.0	64.1
60	96.6	94.9	34.3	30.7	86.9	64.7

Source: US Census Bureau, Survey of Income and Program Participation (SIPP), 1996 Panel, Wave 2 Topical Module

Marriage and Divorce, Table 2 Daily hours of work of men and women (age 20–59) in the market and at home, by marital status, selected countries and years

	US 1985	Can. 1982	UK 1985	Ger. 1992	Italy 1989	Norw. 1990
Paid work						
Single men	5.5	5.6	4.2	6.4	4.9	4.7
Single women	4.6	4.3	3.3	5.0	3.3	4.0
Married men, no child	6.2	6.2	5.5	6.3	5.5	5.7
Married women, no child	3.3	4.0	3.8	3.3	2.0	4.2
Married men, child 5–17	6.1	5.9	5.7	6.7	6.1	6.0
Married women, child 5–17	3.5	3.7	2.6	3.2	2.2	3.6
Married men, child < 5	6.9	6.2	6.1	6.8	6.2	5.7
Married women, child < 5	1.9	2.4	2.0	2.2	1.9	2.1
Housework (*including child care*)						
Single men	1.6	1.7	2.2	1.6	0.7	1.7
Single women	2.8	3.3	3.9	3.4	3.1	2.9
Married men, no child	1.8	2.0	3.3	2.2	1.3	2.1
Married women, no child	4.1	3.9	3.8	4.8	6.4	3.5
Married men, child 5–17	2.3	2.5	2.1	2.3	1.2	2.4
Married women, child 5–17	4.4	4.7	5.5	5.5	7.0	4.5
Married men, child < 5	2.3	3.2	2.3	2.8	1.5	3.2
Married women, child < 5	6.4	6.8	3.8	6.9	7.6	6.1

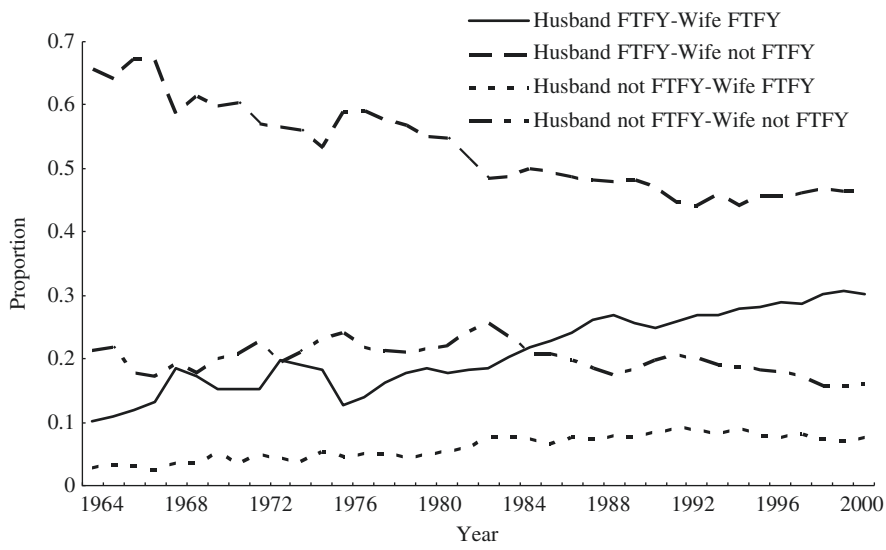
Source: Multinational Time Use Study

old ages. This pattern reflects the earlier marriage of women and their longer lives, which causes the ratio of men to women to decline with age.

One consequence of higher marital turnover is the large number of children who live in single-parent and step-parent households. In 2002, 23 per cent of US children younger than 18 years lived only with their mother, and five per cent lived only with their father. Children of broken families are more likely to live in poverty and to underperform in school. Lower attainments of such children are observed also prior to the occurrence of divorce, suggesting that bad marriage rather than divorce may be the cause (Piketty 2003).

Marriage and Work

Time use data (Table 2) show that men work more than women in the market; women do more housework than men. Per day, single women work at home three hours while single men work less than two hours. These figures roughly double for married couples with young children, showing clearly that children require a substantial investment of time and that most of this load is carried



Marriage and Divorce, Fig. 2 Work patterns of husbands and wives (age 30–40), United States, 1964–2001. *Note*: A spouse is employed full-time-full-year (FTFY) if he/she works 50 weeks or more and hours exceed 34 per week. *Source*: Current Population Surveys

by the mother. The total time worked and the corresponding amount of leisure is about the same for married men and women.

Figure 2 displays the work patterns within couples. The most common situation is that the husband works full-time and the wife works part-time or does not work at all. However, the proportion of such couples has declined and the proportion of couples where *both* partners work full-time has risen sharply, reflecting the increased entry of married women into the labour force.

Marriage and Wages

Male–female wage differences of full-time workers are larger among married than among single persons. Married men have consistently the highest wage among men, while never-married women have the highest wage among women. The wage gap between married men and women rises as the cohort ages, reflecting the cumulative effects of gender differences in the acquisition of labour market experience (Figs. 3 and 4). The increased participation of married women, associated with the increase in their wages, has increased their wage relative to those of never-married women and their husbands (Table 3).

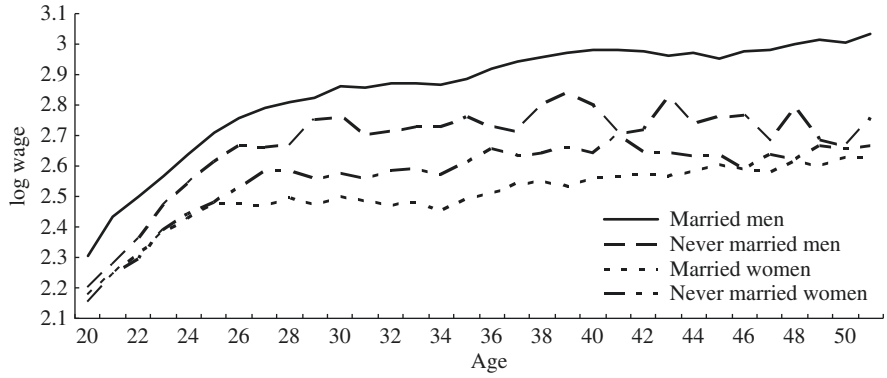
Economic Theory of Marriage and Divorce

From an economic point of view, marriage is a voluntary partnership for the purpose of joint production and joint consumption. As such, it is comparable to other economic organizations that aim to maximize some private gains but are subject to market discipline.

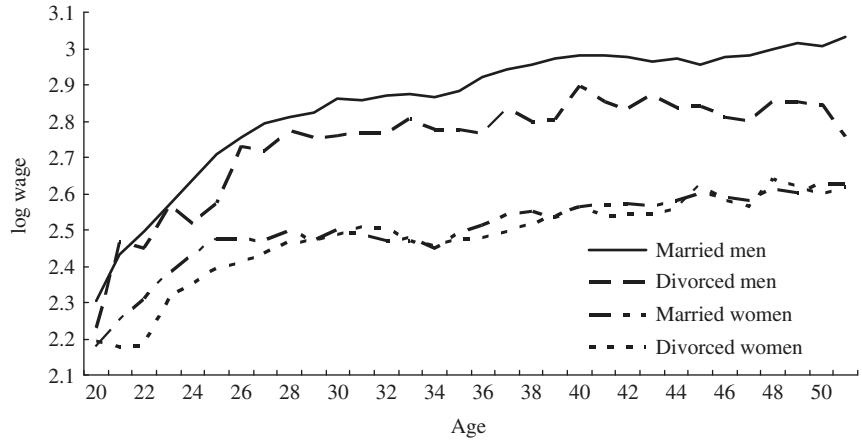
Gains from Marriage

Consumption and production in the family are broadly defined to include non-marketable goods and services, such as companionship and children. Indeed, the production and rearing of children is the most commonly recognized role of the family. We mention here five broad sources of *economic* gain from marriage, that is, why 'two are better than one':

1. Sharing of collective (non-rival) goods; both partners can equally enjoy their children, share the same information and use the same home.
2. Division of labour to exploit comparative advantage or increasing returns; one partner works at home and the other works in the market.



Marriage and Divorce, Fig. 3 Hourly wages (in logs) of fully employed married and never-married US men and women born in 1946–1950, by age. *Source*: Current Population Surveys



Marriage and Divorce, Fig. 4 Hourly wages (in logs) of fully employed married and divorced US men and women born in 1946–1950, by age. *Source*: Current Population Surveys

3. Extending credit and coordination of investment activities; one partner works when the other is in school.
4. Risk-pooling; one partner works when the other is sick or unemployed.
5. Coordination of child care, which is a collective good for the parents. Although children can be produced and raised outside the family, the family has a substantial advantage in carrying out these activities. Two interrelated factors cause this advantage: by nature, parents care about their own children and, because of this mutual interest, it is more efficient that the parents themselves determine the expenditure on their children. If the parents live separately, whether single or remarried, the non-custodian parent loses control of child expenditures. Lack of contact further reduces the incentive or ability to contribute time and money to the children. Together, these factors reduce the welfare of both parents and children when they live apart (Weiss and Willis 1985).

Family Decision Making

The existence of potential gains from marriage is not sufficient to motivate marriage and to sustain it. Prospective mates are concerned whether the potential gains will be realized and how they are divided. Family members have potentially conflicting interests and a basic question is how

Marriage and Divorce, Table 3 Relative wage gaps associated with marital status for fully employed men and women, by year and age, United States, 1965–2001

Years/ age	Married–never married		Married–divorced		Mar. men–mar. women between groups	Husband–wife within couples
	Men	Women	Men	Women		
1965–74						
25–34	13.8	– 8.8	9.6	4.5	37.2	32.5
35–44	21.5	– 17.6	17.1	– 1.6	52.1	42.7
1975–84						
25–34	15.6	– 6.5	8.5	– .5	35.4	29.6
35–44	21.0	– 17.5	12.4	– 2.8	52.1	43.8
1985–94						
25–34	15.6	– 2.0	15.4	7.7	23.6	21.1
35–44	21.3	– 9.9	15.4	2.4	38.7	32.1
1995–2001						
25–34	13.6	2.3	13.6	2.3	17.0	18.1
35–44	23.7	– 1.8	21.4	7.8	31.7	27.5

Source: Current Population Surveys

families reach decisions. The old notion that families maximize a common objective appears to be too narrow. Instead of this *unitary* model, it is now more common to consider *collective* models in which partners with different preferences reach some binding agreement that specifies an *efficient* allocation of resources and a *stable sharing rule*. (Browning et al. 2005, ch. 3).

In a special case, referred to as *transferable utility*, it is possible to separate the issues of efficiency and distribution. This situation arises if there is a commodity (say, money) that, upon changing hands, shifts utilities between the partners at a fixed rate of exchange. In this case, the family decision process can be broken into two steps: actions are first chosen to maximize a weighted *sum* of the individual utilities, and then money is transferred to divide the resulting marital output. In general, the problems of efficiency and distribution are intertwined. We may still describe the family as maximizing a weighted sum of the individual utilities, but the weights depend on the individual bargaining powers, and any shift in the weights will affect the family choice. The bargaining power may depend on individual attributes such as earning capacity, subjective factors such as impatience and risk aversion, and on market conditions, such as the sex ratio and availability of alternative mates (Lundberg and Pollak 1993).

The question remains: what enforces the coordination between family members? One possibility is that the partners sign a formal 'marriage contract' that is enforced by law. However, such contracts are quite rare in modern societies, which can be probably ascribed to a larger reliance than in the past on emotional commitments and the presumption that too much contracting can 'kill love'. In the absence of legal enforcement, efficient contracts may be supported by repeated interactions and the possibility to trade favours and punishments. This possibility arises because marriage is a durable relationship, forged by the long-term investment in children and the accumulation of marital specific capital, which is lost or diminished in value if separation occurs. However, repeated game arguments cannot explain unconditional giving, such as taking care of a spouse stricken by Alzheimers who would never be able to return the favour. Emotional commitments and altruism play a central role in enforcing family contracts (Becker 1991, ch. 8).

The Marriage Market

Individuals in society have many potential partners. An undesired marriage can be avoided or replaced by a better one. This situation creates competition over the potential gains from marriage. In modern societies, explicit price

mechanisms are not observed. Nevertheless, the assignment of partners and the sharing of the gains from marriage can be analysed within a market framework.

Matching models provide a starting point for such analysis. These models investigate the mapping from preferences over prospective matches into a stable assignment (Roth and Sotomayor 1990). An assignment is said to be *stable* if no married person would rather be single and no two (married or unmarried) persons prefer to form a new union. To illustrate, assume that each male is endowed with a single trait, m, and each female is endowed with a single trait, f. Let

$$z = h(m,f). \tag{1}$$

be the *household production function* that summarizes the impact of traits of the matched partners on marital output, z, and assume that $h(m,f)$ is increasing in m and f.

Suppose, first, that z is a public good that the partners must consume jointly. Then, the only stable assignment is such that males with high m marry females with high f, and, if there are more (fewer) eligible men than women, the men (women) with the fewest endowments remain unmarried. All men want to marry the best woman, and she will accept only the best man. After this pair is taken 'out of the game', we can apply the same argument to the next-best couple and proceed sequentially. Such a matching pattern is called positive *assortative matching*.

If one assumes, instead, that z can be divided between the two partners and that utility is transferable, then a man with low m may obtain women with high f by giving up part of his private share in the gains from marriage. The type of interaction in the gains from marriage determines the willingness to pay for the different attributes. Complementarity (substitution) means that the two traits interact in such a way that the benefits from a woman with high f are higher (lower) for a male with high m than for a male with low m. Thus, a positive (negative) assortative matching occurs if the two traits are complements (substitutes). An important lesson is that in a marriage market with sufficient scope for compensation within marriage, the best man is not necessarily the one married to the best women, because, with negative interaction, either one of them can be bid away by the second-best of the opposite sex (Becker 1991, ch. 4).

What determines the division of marital gains? If each couple is considered in isolation then, in principle, any efficient outcome is possible, and one has to use bargaining arguments to determine the allocation. However, in an 'ideal' frictionless case, where partners are free to break marriages and swap partners at will, the outcome depends on the joint distribution of male and female characteristics in the market at large. Traits of the partners in a particular marriage have no direct impact on the shares of the two partners, because these traits are endogenously determined by the requirement of stable matching.

These features show up more clearly if one assumes a continuum of agents and continuous marital attributes. Let $F(m)$ and $G(f)$ be the cumulative distributions of the male and female traits, respectively, and let the measure of women in the total population be r, where the measure of men is normalized to 1. Assume that the female and male traits are complements and transferable utility. Then, if man m' is married to woman f', the set of men with m exceeding m' must have the same measure as the set of women with f above f'. Thus, for all m and f in the set of married couples,

$$1 - F(m) = r(1 - G(f)). \tag{2}$$

This simple relationship determines a positively sloped matching function, $m = (f)$.

A *sharing rule* specifies the shares of the wife and husband in every marriage that forms. Let $v(m)$ be the *reservation utility* that man m requires in *any* marriage and let $u(f)$ be the reservation utility of woman f. Then the sharing rule that supports a stable assignment must satisfy

$$\begin{gathered} v(m) = \max_f \left(h(f,m) - u(f)\right), \\ \text{and} \\ u(f) = \max_y (h(z,y) - v(y)). \end{gathered} \tag{3}$$

That is, each married partner gets the spouse that maximizes his or her 'profit' from the

partnership over all possible alternatives. As we move across matched couples, the welfare of each partner changes according to the *marginal* contribution of his/her *own* trait to the *marital* output, irrespective of the potential impact on the partner whom one marries. With a continuum of agents, there are no rents in the marriage market because everyone receives roughly what can be obtained in the next-best alternative. Another condition for a stable assignment is that, if there are unmarried men, the least attractive married man cannot get any surplus from marriage. Otherwise, slightly less attractive men could bid away his match. A similar condition applies for unmarried women.

From these considerations, one can obtain a unique sharing rule, provided that $r \neq 1$. Basically, one first finds the sharing in the least attractive match, using the no-rent condition. Then the division in better marriages is determined sequentially, by using the condition that along the stable matching profile each partner receives his or her marginal contribution to the marital output. The sharing rule is fully determined by the sex ratio and the respective trait distributions of the two sexes. It can be shown that a marginal increase in the ratio of women to men in the marriage market improves (or leaves unchanged) the welfare of all men, and reduces (or leaves unchanged) the welfare of all women. From (2), it is seen that an upward (downward) first-order shift in the distributions of traits is equivalent (in terms of the effects on the sharing rule) to a marginal increase (decrease) in the female–male ratio. In this regard, there is close correspondence between the impact of changes in quality (that is, the average trait) and size of the two groups that are matched in the marriage market (Browning et al. 2005, ch. 9).

Search

The process of matching in real life is characterized by scarcity of information about potential matches. Models of search add realism to the assignment model because they provide an explicit description of the sorting process that happens in real time.

Following Mortensen (1988), consider infinitely lived agents and assume that meetings are governed by a Poisson random process (these two assumptions are made to ensure a stationary environment). The total marital output is observed upon meeting and, on the assumption of transferable utility, marriage will occur whenever this marital output exceeds the *sum* of the values of continued search of the matched partners. This rule holds because it implies the existence of a division within marriage that makes both partners better off. Because meetings are random and sparse in time, those who actually meet and choose to marry enjoy a positive rent. The division of these rents between the partners is an important issue. Two considerations determine the division of the gains from marriage: outside options, reflected in the value of continued search, and the self-enforcing allocation that would emerge if the marriage continued without agreement (Wolinsky 1987). If these two considerations are combined, the sharing rule is influenced by both the value of search as single and the value of continued search during the bargaining process, including the option of leaving when an outside offer arrives. In this way, a link is created between the division of marital output gains and market conditions.

Search models explain why, despite the gains from marriage, part of the population is not married and individuals move between married and single states. The steady state proportions of the population in each state are such that the flows into and out of each state are equalized. These two flows are determined by the search strategies that individuals adopt.

Search models may have significant externalities. For instance, it may be easier to find a mate if there are many singles searching for mates. There are several possible reasons for such *increasing returns* in the matching process. One reason is that the two sexes meet in a variety of situations (work, sport, social life and so on) but many of these meetings are 'wasted' in the sense that one of the individuals is already attached and not willing to divorce. A second reason is that the establishment of more focused channels, where singles meet only singles, is costly. These will be created only if the 'size of the market' is large enough. Third, the intensity of search by unattached decreases with the proportion of attached

people in the population who are less likely to respond to an offer (Mortensen 1988). In such a case, the marriage (divorce) rates will be above (below) their efficient levels, as each person fails to consider the effect of marriage or separation on the prospects of other participants in the marriage market.

Search and Assortative Matching

The presence of frictions modifies somewhat the results on assortative matching. Following Burdett and Coles (1999), consider a case of non-transferable utility with frictions. Assume that if man m marries women f, he gets f and she gets m. There is a continuum of types with continuous distributions and meetings are generated by a Poisson process with parameter λ. Upon meeting, each partner decides whether to accept the match or to continue the search. Marriage occurs only if both partners accept each other and, by assumption, a match cannot be broken.

Each man (woman) chooses a reservation policy that determines which women (men) to accept. The reservation values for men and women, R_m and R_f, respectively, depend on the individual's own trait. Agents at the top of the distribution of each gender can be choosier because they know that they will be accepted by most people on the other side of the market. Hence, continued search is more valuable for them. Formally, let

$$\begin{aligned} R_m &= b_m + \frac{\lambda\mu_m}{r}\int_{R_m}^{\overline{f}} (f - R_m)dG_m(f), \\ R_f &= b_f + \frac{\lambda\mu_f}{r}\int_{R_f}^{\overline{m}} (f - R_f)dF_f(m) \end{aligned} \tag{4}$$

where the flow of benefits as single, b, the proportion of meetings that end in marriage, μ, and the distribution of 'offers' if marriage occurs, all depend on traits, as indicated by the m and f subscripts. The common discount factor, r, represents the cost of waiting.

In equilibrium, the reservation values of all agents must be a best response against each other, yielding a (stationary) Nash equilibrium. In particular, the 'best' woman and the 'best' man will adopt the policies

$$R_{\overline{m}} = b_{\overline{m}} + \frac{\lambda}{r}\int_{R_{\overline{m}}}^{\overline{f}} (f - R_{\overline{m}})dG(f),$$

$$R_{\overline{f}} = b_{\overline{f}} + \frac{\lambda}{r}\int_{R_{\overline{f}}}^{\overline{m}} \left(m - R_{\overline{f}}\right)dF(m).$$

Thus, the best man accepts some women who are inferior to the best woman and the best woman accepts some men who are inferior to the best man, because a bird in the hand is worth two in the bush.

The assumption that the ranking of men and women is based on a single trait introduces a strong commonality in preferences whereby all men agree on the ranking of all women and vice versa. Because all individuals of the opposite sex accept the best woman and all women accept the best man, μ is set to 1 in Eq. (5) and the distribution of offers equals the distribution of types in the population. Moreover, if the best man accepts all women with f in the range $\left[R_{\overline{m}}, \overline{f}\right]$, then all men who are inferior in quality will also accept such women. But this means that all women in the range $\left[R_{\overline{m}}, \overline{f}\right]$ are sure that all men accept them and therefore will have the *same* reservation value, $R_{\overline{f}}$, which in turn implies that all men in the range $\left[R_{\overline{f}}, \overline{m}\right]$ will have the same reservation value, $R_{\overline{m}}$.

These considerations lead to a *class structure* with a finite number of distinct classes in which individuals marry each other. Having identified the upper class, we can then examine the considerations of the top man and woman in the rest of the population. Lower-class individuals face $\mu < 1$ and a *truncated* distribution of offers because not all meetings end in marriage but, in principle, these can be calculated and then one can find the reservation values for the highest two types and all other individuals in the group forming the second class. Proceeding in this manner to the bottom, it is possible to determine all classes. This pattern is similar to the case without frictions and non-transferable utility except that, because of the need to compromise, low-and high-quality types mix within each class.

With frictions and transferable utility, there is still a tendency towards positive (negative) assortative matching based on the interaction in traits. If the traits are complements, individuals of either

sex with a higher endowment will adopt a more selective reservation policy and will be matched, on the average, with a highly endowed person of the opposite sex. However, with sufficient friction it is possible to have negative assortative matching even under complementarity. This, again, is driven by the need to compromise. With low frequency of meetings and costs of waiting, males with low *m* expect some women with high *f* to accept them. If the gain from such a match is large enough, they will reject all women with low *f* and wait until a high *f* woman arrives.

Divorce and Remarriage

Divorce is motivated by uncertainty and changing circumstances. Thus, individuals may enter a relationship and then break it if a better match is met. Or changing economic and emotional circumstances may dissipate the gains from marriage. As time passes, new information on match quality and outside options is accumulated, and each partner decides whether to dissolve the partnership. In making this choice, partners consider the expected value of each alternative, where the value of remaining married includes the option of later divorce and the value of divorcing includes the option of later remarriage. Under divorce at will, divorce occurs endogenously whenever one partner has an alternative option that the current spouse cannot, or is unwilling to, match by a redistribution of the gains from marriage.

Following divorce, the options for sharing and coordination of activities diminish. The divorced partners may have different economic prospects, especially if children are present. Asymmetries arise because the mother usually loses earning capacity as a result of having a child. To mitigate these risks, the partners have a mutual interest in signing binding contracts that stipulate post-divorce transfers. Such contracts are negotiated 'in the shadow of the law' and are legally binding. Child support payments are mandatory but the non-custodial father may augment the transfer to influence child expenditures by the custodial mother. Payments made to the custodial mother are usually fungible and, therefore, the amount that actually reaches the children depends on the mother's marital status. If she remarries, child expenditures depend on the new husband's net income, including his child-support commitments to his ex-wife. Hence, the willingness of each parent to provide child support depends on commitments of others. These interdependencies can yield *multiple equilibria*, with and without children and correspondingly low and high divorce rates (Browning et al. 2005, ch. 11).

Theory and Evidence

There is a growing body of empirical research that addresses the testable implications of the models outlined above.

1. The unitary model of the household implies that the consumption levels of husband and wife depend only on *total* family income. This, however, is rejected by the data (Lundberg et al. 1997). Nevertheless, consumption and work patterns of married couples indicate that they act efficiently (Browning and Chiappori 1994), implying that a collective model fits the data.
2. Matching models with transferable utility imply positive assortative matching based on the spouses' schooling but negative matching based on their wages (Becker 1991, ch. 10). In fact, the correlation between the education levels of married partners (about.6) is substantially higher than the correlation between their wages (about.3).
3. Because partners are matched based on their traits as observed at the time of marriage, both positive or negative *surprises* trigger divorce (Becker 1991, ch. 10). Weiss and Willis (1997) find an impact of unexpected changes in husband's and wife's incomes on the probability of divorce.
4. Unanticipated shocks are less destabilizing if partners are well matched. Anticipating that, couples would sort into marriage according to characteristics that enhance the stability of marriage. In fact, individuals with similar schooling are less likely to divorce and are more likely to marry. This pattern holds for religion and ethnicity, too (Weiss and Willis 1997).

5. Individual types congregate into locations that facilitate matching; gays in San Francisco (Black et al. 2000) or Jews in New York (Bisin et al. 2004). Such patterns suggest increasing returns in search. Higher wage variability among men induces women to search longer for their first or second husband, consistently with an optimal search strategy (Gould and Paserman 2003).
6. Marital choices and family decisions respond to aggregate marriage market conditions. Black women in the United States delay their marriage and have children out of wedlock because of a shortage of eligible black men (Willis 1999); a higher male–female ratio reduces the hours worked by wives and raises the hours worked by husbands (Chaippori et al. 2002).
7. The sharp increase in divorce in the United States and other countries during 1965–75 seems to constitute a switch across two different equilibria. A marriage market is capable of such abrupt change because of inherent positive feedbacks in matching and contracting. Explanations for the timing of the change include the appearance of the contraceptive pill, the break-up of norms and legal reforms (Michael 1988; Goldin and Katz 2002).

See Also

- ▶ Assortative Matching
- ▶ Collective Models of the Household
- ▶ Marriage Markets

Bibliography

Becker, G. 1991. *Treatise on the family*. Cambridge, MA: Harvard University Press.

Bisin, A., G. Topa, and T. Verdier. 2004. Religious intermarriage and socialization in the US. *Journal of Political Economy* 112: 612–665.

Black, D., G. Gates, S. Sanders, and L. Taylor. 2000. Demographics of the gay and lesbian population in the United States: Evidence from available systematic data sources. *Demography* 37: 139–154.

Browning, M., and P. Chiappori. 1994. Efficient intra-household allocations: Characterization and empirical tests. *Econometrica* 66: 1241–1278.

Browning, M., P. Chiappori, and Y. Weiss. 2005. *Family economics*. Cambridge: Cambridge University Press.

Burdett, K., and M. Coles. 1999. Long-term partnership formation: marriage and employment. *Economic Journal* 109: F307–F334.

Chaippori, P., B. Fortin, and G. Lacroix. 2002. Marriage market, divorce legislation, and household labor supply. *Journal of Political Economy* 110: 37–72.

Goldin, C., and L. Katz. 2002. The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of Political Economy* 110: 730–770.

Gould, E., and D. Paserman. 2003. Waiting for Mr Right: Rising inequality and declining marriage rates. *Journal of Urban Economics* 53: 257–281.

Lundberg, S., and R. Pollak. 1993. Separate spheres bargaining and the marriage market. *Journal of Political Economy* 101: 988–1010.

Lundberg, S., R. Pollak, and T. Wales. 1997. Do husbands and wives pool resources? Evidence from UK Child Benefit. *Journal of Human Resources* 32: 463–480.

Michael, R. 1988. Why did the divorce rate double within a decade? *Research in Population Economics* 6: 367–399.

Mortensen, D. 1988. Matching: Finding a partner for life or otherwise. *American Journal of Sociology* 94-(supplement): s215–s240.

Piketty, T. 2003. The impact of divorce on school performance: evidence from France, 1968–2002. Discussion Paper No. 4146. London: CEPR.

Roth, A., and M. Sotomayor. 1990. *Two sided matching: A study in game-theoretic modeling and analysis*. Cambridge: Cambridge University Press.

Weiss, Y., and R. Willis. 1985. Children as collective goods. *Journal of Labor Economics* 3: 268–292.

Weiss, Y., and R. Willis. 1997. Match quality, new information, and marital dissolution. *Journal of Labor Economics* 15: S293–S329.

Willis, R. 1999. A theory of out-of-wedlock childbearing. *Journal of Political Economy* 107(6): S33–S64.

Wolinsky, A. 1987. Matching, search, and bargaining. *Journal of Economic Theory* 42: 311–333.

Marriage Markets

Andrew Foster

Abstract

The term 'marriage market' refers to the application of economic theory to the analysis of the process that determines how men and women are matched to each other through marriage and how this process influences other choices

including human capital investment and the allocation of marital surplus. The specific sub-topics in this article include a characterization of stable assignment in marriage, consideration of the effects of marriage allocations on distribution within marriage, discussion of the extent to which partners who marry have similar characteristics, and a review of results on marriage timing.

Keywords

Assortative mating; Dowries; Fertility; Human capital investment; Inequality; Marriage markets; Non-market production; Stable assignment; Transferable utility

JEL Classifications

O1

The marriage market is a term used by economists to characterize the process that determines how men and women are matched to each other through marriage. Formally the marriage market may be thought of as an allocative process that, given the preferences and endowments of two sets of individuals (men and women), yields a set of couples and unmatched individuals and a distribution of resources within each match. Marriage markets are generally distinguished from other sorting processes such as worker–firm matching by the assumption that each member of each set of individuals is matched to at most one member of the other set. However, the basic concept of the marriage market may also be applied to other cases such as polygamy or same-sex partnerships. It is also generally assumed that one's well-being within marriage is determined by the characteristics of one's partner and the distribution of resources within the marriage, but not the matches of other individuals in the marriage market conditional on these factors. The economics literature on the marriage market has built importantly on a two-part foundational article on the economics of marriage published in 1973 and 1974 (Becker, 1973, 1974). However, the phrase 'marriage market' is considerably older, with a first citation in the *Oxford English Dictionary* of 1842.

Stable Assignment

Central to the notion of a marriage market is the notion of stable assignment. A stable assignment may be characterized as a set of partner allocations and distributions of resources within marriage so that no individual of one sex would be willing to make an offer (in terms of partnership and a distribution of resources within that partnership) to an individual of the other sex which that individual strictly prefers to his or her equilibrium allocation.

An early and important divide in terms of economic models of marriage arises with respect to the question of transferable utility. Transferable utility arises when well-being within the household may be freely transferred between members of the household through a reallocation of household resources. Under these conditions the question of who marries whom can be importantly separated from the question of how resources are distributed within marriage and any stable marriage assignment can be characterized as the outcome of the maximization of a linear programme (Bergstrom, 1997).

At the other extreme from a transferable utility model is one in which there is no possibility of transferring resources within or across marriage. A key feature of such models is that there is generally a wide variety of possible stable equilibria. Gale and Shapley (1962) illustrate two such stable equilibria, by the construction of two matching algorithms based on who makes offers and who makes the decision to accept, tentatively accept, or reject those offers. Each man is at least as well off in the equilibrium in which men make offers relative to the equilibrium in which women makes offers and vice versa.

Distributive Effects

Becker's (1973) pioneering analysis of the marriage market considered, among other things, the effects of the marriage market on household distribution. Consider, for example, a simplified version of this model in which there is heterogeneity in tastes for being single, transferable utility

within marriage, and no heterogeneity across couples in total utility within marriage. The outcome of the model is a distribution parameter that characterizes the share of total marital utility going to each partner within marriage and a number of marriages, with those individuals of both sexes with the highest taste for being single remaining unmarried. Among other things the model illustrates how a rise in the female wage raises the utility of married females within marriage even when married women are not active in the labour market. The increased opportunities for women outside marriage implies that women must, at the margin, receive a higher share of marital utility in order to be willing to marry.

There is substantial debate about the importance of marriage market structure in influencing transfers between partners and their respective households of origin at the time of marriage. Of particular relevance is the evidence of a historical transition from bride-price to dowry in parts of South Asia and the very large levels of dowry relative to annual income that are sometimes observed in that region. A number of factors have been argued to play an important role in this regard, including changes in the relative sizes of female and male populations of marriageable age associated with population growth and the gap in typical ages at marriage, changes in inequality and economic opportunity, and changes in the relative merits of different forms of parental transfers in their children.

Assortative Mating

A second issue that has received significant theoretical and empirical scrutiny is the question of the extent to which the marriage market matches men and women with similar characteristics. This issue is thought to be important because of its implications for interhousehold inequality and for intergenerational transmission of inequality. If high-earning men match with high-earning women, and these high-earning couples transfer these resources to their children in the form of financial assistance and/or human capital, then inequality is likely to be more persistent across generations than would be the case otherwise. Assortative mating by religion and/or immigrant status is also thought to be both an indicator of and contributor to the process of assimilation. Finally, assortative mating on unobservable (to analysts) attributes can affect inferences about household behaviour that condition on household composition. For example, if men with a high unobservable taste for child human capital match with more educated women, then highly educated women will appear to have more educated children even if there is no direct effect.

A simple transferable utility model in which marital output is increasing in the product of male and female quality yields the prediction that there should be positive assortative mating on such attributes as intelligence, wealth and beauty. A possible exception arises with respect to market earnings capacity to the extent that, as postulated by Becker (1973), one member of the couple specializes in the production of non-market goods. Interestingly, the theoretical prediction of positive assortative mating across classes of individuals can arise within the marriage market with imperfect information (Burdett and Coles, 1997).

The evidence supports the prediction of positive assertive mating on partner attributes, although there have been changes over time in the degree to which this is observed. In particular, the degree of educational assortative mating fell between 1940 and 1960 in the United States but has increased subsequently, largely due to a decline in the share of low-education individuals marrying (Schwartz and Mare, 2005). There has also been a shift in the sign of the correlation in partner earnings from negative to positive since the 1960s (Schwartz, 2005), a pattern that has contributed to the overall increase in interhousehold inequality in income.

Marriage Timing

A third set of marriage-market issues relates to the timing of marriage, particularly for women. It is argued that early marriage can result in higher fertility, lower rates of human capital investment, and an adverse bargaining position from the

perspective of women. Boulier and Rosenzweig (1984), in an early contribution on this subject, showed how unobserved attractiveness could lead to incorrect inference about the role of education in delaying marriage and increasing spousal quality. Bergstrom and Bagnoli (1993) show how the process of uncertainty resolution with regard to the marital prospects may differentially affect the timing of marriage for men and women of different qualities. It also is the case that timing of marriage can play an important role in the equilibration of marriage markets given substantial differences in the relative numbers of eligible men and women arising from sex differences in mortality or a gap in the age at marriage for men and women for a growing population. In particular, because of how changes in the timing of marriage for sequential cohorts of eligible men and women affect the number of marriages taking place at a particular point in time, a persistent ten per cent excess in the number of eligible females relative to males can be accommodated with an increase in the female relative to male age at marriage by just one year over a decade (Foster et al. 2004).

See Also

- ▶ Assortative matching
- ▶ Becker, Gary S. (Born 1930)
- ▶ Family economics
- ▶ Household production and public goods
- ▶ Marriage and divorce
- ▶ Matching and market design

Bibliography

Becker, G.S. 1973. A theory of marriage: Part I. *Journal of Political Economy* 81: 813–846.

Becker, G.S. 1974. A theory of marriage: Part II. *Journal of Political Economy* 82: S11–S26.

Bergstrom, T. 1997. A survey of theories of the family. In *Handbook of population and family economics*, ed. M.-R. Rosenzweig and O. Stark, Vol. 1A. New York: North-Holland.

Bergstrom, T.C., and M. Bagnoli. 1993. Courtship as a waiting game. *Journal of Political Economy* 101: 185–202.

Boulier, B., and M. Rosenzweig. 1984. Schooling, search and spouse selection: Testing economic theories of marriage and household behavior. *Journal of Political Economy* 92: 712–732.

Burdett, K., and M. Coles. 1997. Marriage and class. *Quarterly Journal of Economics* 112: 141–168.

Foster, A., N. Khan, and A. Protik. 2004. Equilibrating the marriage market in a rapidly growing population: Evidence from rural Bangladesh. Working paper, Department of Economics, Brown University.

Gale, D., and L. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Schwartz, C.R. 2005. Earnings inequality among married couples and the increasing association between spouses' earnings. Working paper, Department of Economics, UCLA.

Schwartz, C.R., and R.D. Mare. 2005. Trends in educational assortative marriage from 1940 to 2003. *Demography* 42: 621–646.

Marschak, Jacob (1898–1977)

Roy Radner

M

Keywords

Decision theory; Demand theory; Econometrics; Equilibrium; Information, value of; Joint demand; Marschak, J.; Rationality; Socialism; Stochastic theory; Teams, economic theory of; Uncertainty

JEL Classifications

B31

The diversity of Jacob Marschak's education and early experience made it likely that he would approach the study of economic behaviour with more than the average breadth of interest and vision. He was born in Kiev on 23 July 1898, and studied mechanical engineering at the Kiev Institute of Technology. At the beginning of the Russian Revolution he served briefly as Minister of Labour in the Menshevik government of Georgia but was forced to escape to Germany. There he went first to the University of Berlin, where he studied economics and statistics with

L.V. Bortkiewicz, and then to the University of Heidelberg, where he received his Ph.D. in economics in 1922. His professors at Heidelberg included E. Lederer in economics, A. Weber in sociology, K. Jaspers in philosophy and G. Anschuetz in public law.

Following his doctoral studies at Heidelberg, he earned his living for the next eight years as an economic journalist and applied economist. He was economic editor for the *Frankfurter Zeitung* (1924–5), a research associate at the Research Centre for Economic Policy in Berlin (1926–8), and supervisor and editor of research for a Parliamentary Commission of Exporting Industries, at the Institute of World Economics of the University of Kiel (1928–30). Also, in 1926 he spent time in London on a travelling fellowship from the University of Heidelberg.

In 1930 he was appointed as a Privatdozent in economics at the University of Heidelberg, but three years later, once again the victim of political events, he left Germany and went to Oxford as a university lecturer. In 1935 he became Reader in Statistics and Director of the Oxford Institute of Statistics, where he remained until 1939. During this period he wrote extensively on theoretical and statistical aspects of demand analysis, a field in which he was a pioneer (Marschak 1931).

In 1939 Marschak moved to the United States, where he lived the rest of his life, teaching at the New School for Social Research (1940–42), the University of Chicago (1943–55), Yale University (1955–60), and the University of California at Los Angeles (1960–77).

During the first dozen years Marschak was an active participant in the econometric revolution that is commonly associated with the Cowles Commission for Research in Economics. This revolution was nurtured at an early and crucial stage by the seminar on econometric methods and results that Marschak organized at the National Bureau of Economic Research, while he was on the faculty of the New School for Social Research. The intensive contacts fostered in this seminar led, in particular, to three fundamental papers on the statistical estimation of systems of simultaneous equations, by Haavelmo (1943), Mann and Wald (1943), and Marschak and Andrews (1944). Two further publication landmarks in this movement were the Cowles Commission Monographs No. 10 and No. 14, to which Marschak contributed the opening chapters (Marschak 1950a, 1953).

Two other topics on which Marschak worked presaged his later work on decision and organization. First, he was, for a number of years, interested in the demand for money, and through his work and that of others the idea evolved that this demand could be better understood in the context of a more general theory of the joint demand for various assets (Marschak 1938, 1949, 1950b). Furthermore, since the ultimate values of assets are rarely known with certainty at the time they are acquired, such a general theory needed to be based on a more systematic theory of decision in the face of uncertainty than was then available.

A second topic was the subject of his first scientific publication, a contribution to the debate on the efficiency, or even viability, of socialism. A central issue in that debate was whether the centralization of economic authority in a socialist state was compatible with the decentralization of information necessary in a complex economy.

From 1950 on, Marschak's research and writing was concerned with the general area of decision, information and organization. More specifically, one can identify at least three topics to which he made substantial contributions: (1) stochastic decision, (2) the economic value of information, and (3) the theory of teams.

Stochastic Decision

In a series of articles (Marschak 1959a, 1964a; Marschak and Block 1960; Marschak and Davidson 1959b; Marschak et al. 1963a, b, 1963c, 1964b), Marschak proposed and elaborated the theory of stochastic decision and reported on a number of experiments. This work had its roots in the theory of rational economic choice or utility theory and in certain theories of psychological measurement.

Marschak developed a framework for describing the behaviour of economic decision makers who are approximately rational or consistent, or whose consistency of behaviour cannot be exactly verified through observation because of the observer's inability to control or identify all of the relevant factors in the decision-making situation.

It had long been recognized that economic decision makers did not exhibit exact consistency in their detailed choices. Economists were and remain loath to abandon the general framework of rational decision making that has appeared to be so fruitful in the analysis of the economic system as a whole, Marschak's theory provided a theoretical model that could be used for econometric studies of individual choice behaviour and that was connected in a coherent way with the general hypothesis of economic rationality. The work of Marschak and his co-authors was at first more appreciated by psychologists than by economists. His papers on this subject are still standard references in the theory of psychological scaling (Luce et al. 1963, vol. 3, ch. 19). More recently, this theory has provided the basis of statistical studies of individual choice behaviour (McFadden 1982), as well as of a new approach to the theory of economic equilibrium that takes account of the uncertainty of individual behaviour (Hildenbrand 1971; Bhattacharya and Majumdar 1973).

Economic Value of Information

Marschak was probably the first to develop a systematic theory of the economic value of information. In this development he recognized that the measurement of quantity of information used by communication engineers, and associated with the work of Wiener and Shannon, was not adequate to measure the value of information. Indeed, it was not possible to identify a single measure of information such that more is always better.

Instead, Marschak turned to the newly developed theory of statistical decision for the source of his framework. For him, the value of a particular information system – or more generally, a system of information gathering, communication and decision – was related to the particular class of economic decision problems under consideration. His theoretical analysis of the value and cost of information pointed to the importance of more empirical knowledge concerning the technology of observation, information processing, communication and decision making, although he, himself, did not do any empirical work in this field. These ideas are elaborated in a long series of papers beginning with his contribution to *Decision Processes* (Marschak 1954) and summarized in his paper 'Economics of Information Systems' (1971).

Economic Theory of Teams and Organization

In an economic or other organization, the members of the organization typically differ in (1) the actions or strategies available to them, (2) the information on which their actions can be based, and (3) their preferences among alternative outcomes and their beliefs concerning the likelihoods of alternative outcomes given any particular organization action. Marschak recognized that the difficulty of determining a solution concept in the theory of games was related to differences of type 3. However, a model of an organization in which only differences of types 1 and 2 existed, which he called a team, presented no such difficulty of solution concept, and promised to provide a useful tool for the analysis of problems of efficient use of information in organizations. Such a model provided a framework for analysing the problems of decentralization of information so central to both the theory of competition and the operation of a socialist economy. The idea of a team was introduced in Marschak (1954, 1955), and a systematic development of the theory of teams is provided in Marschak and Radner (1972).

Towards the end of his career, Marschak returned to the theoretical issues concerning conflict of interest among the members of a decentralized organization. He approached this primarily in terms of the normative problem of

devising incentives for the members of a 'team' to behave in accord with the goals of the organization. Of course, to the extent that such incentives are needed, the organization is no longer a team, in the technical sense of the term and the problem is back in the domain of the more general theory of games. It was left to others to make substantial progress on this set of problems. An important early effort in this direction was by T. Groves, who in his doctoral dissertation (1969) and his subsequent article, 'Incentives in Teams' (1973) presented – in a particular case – a solution to the problem of providing incentives to decentralized decision makers to both send truthful messages and make optimal decisions. These ideas were further developed in the contexts of the theory of public goods, the allocation of resources in a divisionalized firm and the principal–agent relationship. (For references to the literature on these developments see Groves and Ledyard 1987; Hurwicz 1979; Radner 1986.)

Besides the significance of Marschak's individual contributions to economic analysis, I would like to emphasize the cumulative significance of his life's work. Through his work ran the important message that economists must come to grips with problems of uncertainty. He led the way, not only through his own research, but through his indefatigable and successful efforts at explaining these problems to his colleagues in economics and related disciplines. His work drew from psychology, statistics and engineering, and in turn influenced research in those disciplines. Indeed, more than any other economist I know, Marschak typified the best in behavioural science.

Selected Works

A bibliography of Marschak's publications (excluding most book reviews and all newspaper articles) can be found in McGuire and Radner (1986). A large number of his papers have been reprinted in Marschak (1974).

1923. Wirtschaftsrechnung und Gemeinwirtschaft. *Archiv für Sozialwissenschaft* 51: 501–520.

1931. *Elastizität der Nachfrage*. Tübingen: J.C.B. Mohr.

1938. Money and the theory of assets. *Econometrica* 6: 311–325.

1944. (With W.H. Andrews.) Random simultaneous equations and the theory of production. *Econometrica* 12: 143–205.

1949. Role of liquidity under complete and incomplete information. *American Economic Review* 39: 182–195.

1950a. Statistical inference in economics: An introduction. In *Statistical inference in dynamic economic models*, ed. T.C. Koopmans. New York: Wiley.

1950b. The rationale of the demand for money and 'money illusion'. *Metroeconomica* 2: 71–100.

1953. Economic measurements for policy and prediction. In *Studies in econometric method*, ed. W.C. Hood and T.C. Koopmans. New York: Wiley.

1954. Towards an economic theory of organization and information. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis. New York: Wiley.

1955. Elements for a theory of teams. *Management Science* 1: 127–137.

1959a. Binary-choice constraints and random utility indicators. In *Mathematical methods in social sciences*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

1959b. (With D. Davidson.) Experimental tests of stochastic decision theory. In *Measurement: Definitions and theory*, ed. C.W. Churchman and P. Ratoosh. New York: Wiley.

1960. (With H.D. Block.) Random orderings and stochastic theories of responses. In *Contributions to probability and statistics*: Essays in honor of Harold Hotelling, ed. I. Olkin et al. Stanford: Stanford University Press.

1963a. (With G. Becker and M. DeGroot.) Stochastic models of choice behavior. *Behavioral Science* 8: 41–55.

1963b. (With G. Becker and M. DeGroot.) An experimental study of some stochastic models for wagers. *Behavioral Science* 8: 199–202.

1963c. (With G. Becker and M. DeGroot.) Probability of choices among very similar objects: an experiment to decide between two models. *Behavioral Science* 8: 306–311.
1964a. Actual versus consistent decision behavior. *Behavioral Science* 9: 103–110.
1964b. (With G. Becker and M. DeGroot.) Measuring utility by a single-response sequential method. *Behavioral Science* 9: 226–232.
1971. Economics of information systems. In *Frontiers of Quantitative Economics*, ed. M. Intriligator. Amsterdam: North-Holland.
1972. (With R. Radner.) *Economic theory of teams*. New Haven: Yale University Press.
1974. *Economic information*, *decision*, *and prediction*. 3 vols. Dordrecht: Reidl.

Bibliography

Bhattacharya, R.N., and M.K. Majumdar. 1973. Random exchange economies. *Journal of Economic Theory* 6: 37–67.
Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.
Groves, T., and J. Ledyard. 1987. Incentive compatibility ten years later. In *Information, incentives, and economic mechanisms: Essays in honor of Leonid Hurwicz*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.
Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11: 1–12.
Hildenbrand, W. 1971. Random preferences and equilibrium. *Journal of Economic Theory* 3: 414–429.
Hurwicz, L. 1979. On the interaction between information and incentives in organizations. In *Communication and control in society*, ed. K. Drittendorf. New York: Gordon Breach.
Luce, R.D., R. Bush, and E. Galanter. (eds.). 1963–5. *Handbook of mathematical psychology.* 3 vols. New York: Wiley.
Mann, H.B., and A. Wald. 1943. On the statistical treatment of linear stochastic difference equations. *Econometrica* 11: 173–220.
McFadden, D. 1982. Qualitative choice models. In *Advances in economic theory*, ed. W. Hildenbrand. Cambridge: Cambridge University Press.
McGuire, C.B., and R. Radner (eds.). 1986. *Decision and organization*, 2nd ed. Minneapolis: University of Minnesota Press. Originally published Amsterdam: North-Holland, 1972.
Radner, R. 1986. The internal economy of large firms. *Economic Journal* 96: 1–22.

Marshall Plan

Francisco Alvarez-Cuadrado

Abstract

The Marshall Plan transferred over US$12.5 billion to Western European countries between 1948 and 1951. This article contrasts the main views on its impact on the post-war European performance. It concludes that, although the direct impact of the plan through private and public investment was rather limited, Marshall Aid provided the recipient economies with a temporary solution for the severe dollar constraint that posed a threat to the continuation of the European miracle. Furthermore the Plan played an important role in promoting collaboration among former adversaries.

Keywords

Aid; Marshall Plan; OEEC; Post-war economics

JEL Classifications

N14; F35

In Europe during the Second World War the productive effort of more than an entire generation was lost, with per capita income returning to the levels of the turn of the century. This fall in output reflected not only the destruction of capacity but also the disruption of channels for obtaining inputs and distributing production. The reconstruction process began right after the war, with industrial production reaching pre-war levels as soon as 1947. The weather conditions in 1947 substantially depressed agricultural yields, leading to important food and energy shortages. The substantial trade deficits of the recovering economies combined with the negative experience of international investors after the First World War lead to a 'dollar gap' which posed a threat to the continuation of the European miracle.

These were the circumstances that surrounded the development of the Marshall Plan, officially the European Recovery Program (ERP). The views on the motivations behind the ERP range from plain American imperialism (Kolko and Kolko 1972) to pure altruism; in the words of Galbraith (1998), 'the primary purpose of the Plan was compassionate good will, the notion that our former allies needed to have the help of the US'. Nonetheless most of the literature acknowledges that, beyond the concern for former allies, the political and social stability of non-communist Europe and the continuity of export markets for US products were two of the main concerns of the Truman administration.

The initial Plan proposal was presented in June 1947 by secretary of state George Marshall, asking European governments to design a coordinated aid programme to be funded by the United States. The offer included the Soviet Union and its allies, but the conditional terms on economic collaboration and disclosure of information guaranteed that the Soviet Union would never accept it. In response to the American offer the final aid recipients, Austria, Belgium, Denmark, France, West Germany, Great Britain, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Sweden, Switzerland and Turkey, formed the Organization for European Economic Cooperation (OEEC) to coordinate a proposal based on national needs and consistent with American objectives on trade and economic cooperation between recipient countries. US President Truman signed the Plan into a law on 3 April 948, establishing the American-led Economic Cooperation Administration (ECA) to administer the programme.

Over the four years that followed its approval by Congress, the Plan transferred $12.5 billion of US aid to Western Europe. The allocation of funds did not follow a simple rule, although it was mainly determined by the dollar balance of payment deficits of the recipient economies, taking also into account geopolitical considerations especially in the cases of France and the United Kingdom. Marshall Aid represented 2.1 per cent of US GNP in 1948, rose to 2.4 per cent in 1949 and then fell to 1.5 per cent in the remaining two years. In terms of national income of the recipient economies, the funds ranged from 0.3 per cent per year for Sweden to 14 per cent for Austria. For the large Western European economies it represented an average yearly transfer of 2.5 per cent of GDP for France, 2.2 per cent for Italy, 1.3 per cent for the United Kingdom and 1.2 per cent for Germany.

The OEEC took the leading role in allocating funds, conditional on the approval of the ECA. The American supplier was paid in US dollars, which were debited against the ERP account corresponding to the European buyer. This buyer paid for the American imports in local currency, which was deposited by its own government in a counterpart fund. These additional resources were used for local investment projects and eventually were absorbed into the recipient's national budget.

The impact of the plan on the European recovery is not free of controversy. On the one hand, early triumphalist accounts (Jones 1955; Mayne 1970; Arkes 1972) describe the Plan as vital for the reconstruction of productive capacity, the development of the necessary institutions for cooperation among former adversaries, and the restoration of European confidence in market capitalism. In the words of Mayne (1970), Marshall Aid 'was a precondition of all later affluence and economic miracles, as well as moves toward European unity'. On the other hand, Milward (1984) discounts the importance of ERP transfers, arguing that the recovery was well under way before 1948 and the reconstruction of the damaged private and public capital stocks was almost completed. Somewhere in between, De Long and Eichengreen (1993) argue that Marshall Aid helped the recipient economies more in terms of political economy than macroeconomics. The ERP bought European governments the political space needed to avoid the attrition wars that characterized the interwar period, allowing an institutional environment conducive to growth.

Until the influential work of Milward (1984), the literature agreed on the vital importance of the ERP funds for Western European growth. According to this view Marshall Aid allowed for the reconstruction of the capital stock, the elimination of bottlenecks that obstructed production,

the public provision of infrastructures and the surge in intra-European trade. In the words of Arkes (1972), 'the plan was critical at the margins having a multiplier effect of three or four times its value'. A superficial analysis of the data suggests that this view is exaggerated. If we compute the growth rates of per capita GDP for the recipient economies, we find that in the three years that preceded the ERP, yearly growth averaged 6.5 per cent, while during the Plan it averaged 4.4 per cent, falling to four per cent between 1953 and 1956.

Eichengreen et al. (1992) argue that if the effects of the Plan worked mainly through private and public capital accumulation, there should be a significant correlation between output growth and ERP allotments as a share of GDP across recipient economies. Contrary to this view, their statistical analysis does not turn up a significant coefficient on Aid allotments. Alvarez-Cuadrado and Pintea (2008) present a two-sector neoclassical growth model with public capital to explore the direct impact of the ERP. Their numerical analysis suggests that the transfers increased the rate of private investment by less than one percentage point, leading to no more than half a percentage point increase in the growth rate of output. Since average yearly transfers represented no more than half a year's worth of post-war growth, it is not surprising that the effects of the Plan through capital accumulation are rather limited. Along similar lines, Milward (1984) argues that the outstanding performance of Western European economies would have not been very different in the absence of the Plan. He discounts the direct impact of the Plan and convincingly documents that the leverage afforded by the ERP was insufficient for the United States of America to force through its vision of the United States of Europe. In Milward's view the primary role of the ERP was limited to sustaining the flow of capital imports necessary to prolong the recovery.

In a different spirit, De Long and Eichengreen (1993) argue that political economy considerations lie behind the true impact of the Plan. Marshall Aid provided the currency needed to relax the foreign exchange constraint, giving European policy makers extra room to manoeuvre. This political space, together with aid conditionality, induced European governments to balance their budgets, restore internal financial stability, and maintain their commitment to free markets. Their counterfactual vision of Western Europe suggests a permanent influence of Communist parties, an expansion of government controls and regulations, and a resurgence of economic nationalism and isolationism.

This argument, although it has its merits, does not seem to account for the variety of institutional arrangements present in the recipient economies. For instance, two of the fastest-growing economies, France and Germany, adopted rather different growth strategies. The French economy was characterized by major involvement of the state in key economic sectors, while the German approach illustrates the growth potential of relatively free markets. In my view, although the functioning of the market mechanism was constrained in many countries as a result of war priorities, there was a tacit consensus that this was only a temporary interruption of the long European experience with free markets, and therefore the influence of the Plan was rather limited in this respect.

To sum up, the direct impact of the Plan led to no more than half of a percentage point of growth per year. Along the lines of Milward (1984), I believe the plan provided European governments the means to prolong the recovery process which began after the war. In some cases the transfers complemented export revenues, preventing a balance-of-payment crisis, while in others they only postponed its occurrence. The political economy argument is more difficult to evaluate, but given the prior European experience with free markets and the existence of a well-developed system of property rights, it is difficult fully to accept the counterfactual scenario drawn by De Long and Eichengreen (1993). Finally, the Marshall Plan played an important role by inducing British and French support for a strong Germany. Although the forces that led to the process of European integration responded more to internal political and economic developments in the European countries than to American pressure, the Marshall Plan helped promote collaboration among former adversaries.

See Also

► Foreign Aid

Acknowledgment I would like to thank Markus Poschke for helpful comments and suggestions. Of course all remaining errors are mine.

Bibliography

Alvarez-Cuadrado, F., and M. Pintea. 2008. A quantitative exploration of the golden age of European growth. Florida International University, Department of Economics Working Paper no. 0805 (forthcoming in *Journal of Economic Dynamics and Control*).

Arkes, H. 1972. *Bureaucracy, the Marshall Plan, and the national interest*. Princeton: Princeton University Press.

De Long, J.B., and B. Eichengreen. 1993. The Marshall Plan: History's most successful structural adjustment program. In *Postwar economic reconstruction and lessons for the East today*, ed. R. Dornbusch, W. Nolling, and R. Layard. Boston: MIT Press.

Eichengreen, B., M. Uzan, N. Crafts, and M. Hellwig. 1992. The Marshall Plan: Economic effects and implications for Eastern Europe. *Economic Policy* 7(14): 14–75.

Galbraith, J.K. 1998. Testimony in *The Cold War*, directed by T. Coombs.

Jones, J.M. 1955. *The fifteen weeks: An inside account of the genesis of the Marshall Plan*. New York: Viking.

Kolko, J., and G. Kolko. 1972. *The limits of power: The World and United States foreign policy 1945–54*. New York: Harper & Row.

Mayne, R. 1970. *The recovery of Europe*. Garden City: Anchor Press.

Milward, A.S. 1984. *The reconstruction of Western Europe 1945–51*. London: Methuen.

Marshall, Alfred (1842–1924)

John K. Whitaker

Abstract

English economist Alfred Marshall, founder of the Cambridge School of economics, was a leading and internationally prominent figure in the development of economic thought between 1870 and 1920. He played a significant role in professionalizing British economics, always stressing the social importance of wider economic understanding. His influential ideas on economic theory were conveyed primarily in his *Principles of Economics* (1890). Accounts of his life, career and general views on economics are followed by a more technical treatment of his contributions to various aspects of economic analysis. Guides to his writings and to the secondary literature on him are appended.

Keywords

Bernoulli hypothesis; Berry, A.; Bimetallism; Böhm-Bawerk, E.; Bowley, A.; British classical economics; British Economic Association; Cambridge School; Ceteris paribus; Chapman, S.; Clapham, J.; Clark, J.; Combinations; Comparative static method; Conspicuous consumption; Consumer surplus; Cost and supply curves; Cost of production; Cournot, A.; Cunynghame, H.; Cyclical unemployment; Deductive method; Demand for money; Demand price; Demand theory; Derived demand; Differential rent; Distribution theories (classical); Dupuit, J.; Edgeworth, F.; Envelope theorem; Equilibrium analysis; Excess burden of taxation; External economies; Fay, C.; Flux, A.; Foxwell, H.; Free trade; Gonner, E.; Homo economicus; Imperfect competition; Institutions; Internal economies; International trade (theory); Interpersonal utility comparisons; Jenkin, H.; Jevons, W.; Joint demand and supply; Keynes, J. M.; Keynes, J. N.; Lavington, F.; Long-period equilibrium; Long-period supply; Macgregor, D.; Marginal cost pricing; Marginal productivity theory; Marginal utility of money; Marginal utility theory; Market demand elasticity; Market interdependence; Market value; Marshall, A.; Marshall, M. P.; Menger, C.; Methodenstreit; Methodology of economics; Mill, J. S.; Monetary theory; Monopolistic competition; Monopoly; Multiple equilibria; Multiplier; Net product; Nicholson, J.; Normal profit; Normal value; Offer curves; Optimal taxation; Pantaleoni, M.;

Pareto, V.; Partial-equilibrium method; Perfect competition; Period analysis; Pigou, A.; Poverty; Price, L.; Principle of substitution; Producer surplus; Product differentiation; Quasi-rent; Ramsey price; Rent; Representative firm; Ricardo, D.; Royal Economic Society; Sanger, C.; Say's Law; Self-interest; Short-period equilibrium; Sidgwick, H.; Smith, A.; Subjective theories of value; Supply and demand; Tariffs; Temporary equilibrium; Trade unions; Utilitarianism; von Thunen, J.; Walker, F.; Walras, L.; Welfare economics

JEL Classifications
B31

Alfred Marshall, Professor of Political Economy at the University of Cambridge from 1885 to 1908 and founder of the Cambridge School of Economics, was born in Bermondsey, a London suburb, on 26 July 1842. He died at Balliol Croft, his Cambridge home of many years, on 13 July 1924 at the age of 81. His magnum opus, *Principles of Economics* (1890a) evolved through eight editions in his lifetime, the final edition (1920) being most commonly cited today. It was one of the most influential treatises of its era and was for many years the Bible of British economics, introducing many still familiar concepts. The Cambridge School rose to great eminence in the 1920s and 1930s. A.C. Pigou and J.M. Keynes, the most important figures in this development, were among Marshall's pupils.

Marshall's biography and career are outlined initially, after which descriptions are given of his views on the social setting, aims and methods of economics, and his intellectual debts to others. An analysis of his fundamental ideas on theories of value and distribution, which were mainly set out in *Principles*, follows, after which his contributions to monetary and international-trade theory are considered briefly. A final section provides additional documentation and general suggestions for further reading. Also, some of the more technical sections have attached to them brief 'bibliographic notes' offering suggestions for further exploration. All bibliographic references lacking an author's name are to works by Marshall, and the bibliographic details of all his cited publications can be found in the list of 'Selected works' below. The bibliographic details for all cited works written or edited by others are listed in the concluding 'Bibliography'.

Biography and Career

Marshall grew up in the London suburb of Clapham, being educated at the Merchant Taylors' School where he showed academic promise and a particular aptitude for mathematics. Eschewing the more obvious path of a closed scholarship to Oxford and a classical education, he entered St John's College, Cambridge, in 1862 on an open exhibition. There he read for the Mathematical Tripos, Cambridge University's most prestigious degree competition, emerging in 1865 in the exalted position of Second Wrangler, bettered only by the future Lord Rayleigh. This success ensured Marshall's election to a Fellowship at St John's. Supplementing his stipend by some mathematical coaching, and abandoning – doubtless because of a loss of religious conviction – half-formed earlier intentions of a clerical career, he became engrossed in the study of the philosophical foundations and moral bases for human behaviour and social organization. In 1868 he became a College Lecturer in Moral Sciences at St John's, coming to specialize in teaching political economy. By about 1870 he seems to have committed his career to developing this subject, seemingly ripe for reform, and helping to transform it into a new science of economics.

For several years he laboured persistently to develop and refine his economic ideas, and to deepen his understanding and grasp of both the existing economic literature and the economic reality that was its subject matter. In 1875 he visited the United States to probe economic conditions, and throughout his life he was tireless in his efforts to master the practicalities of the economic world. Prior to 1879 his publications were meagre. He had embarked on a book on international trade and problems of protectionism in the

mid-1870s, and before that he had worked out many of his distinctive theoretical ideas in the form of short essays, many now reproduced in Whitaker (1975). But the only part of this material to be made public was four chapters from the theoretical appendices for the proposed international-trade volume. In 1879 Henry Sidgwick had these printed for private circulation under the title *The Pure Theory of Foreign Trade: The Pure Theory of Domestic Values* (1879a). (An amplified version together with surviving portions of the text of the abandoned trade volume is also reproduced in Whitaker 1975.) The year 1879 also saw the publication of Marshall's first book, *The Economics of Industry* (1879b), written jointly with his wife Mary Paley Marshall.

Mary Paley had been one of the first group of students at Newnham Hall (later Newnham College) where Marshall, an early supporter of the informal scheme of Cambridge lectures for women, taught her political economy. Their marriage in 1877 required Marshall to give up his Cambridge position under the celibacy rules then in force. He found a new livelihood as principal of the recently established University College, Bristol, where he also became Professor of Political Economy. There *The Economics of Industry* was brought to completion and published by the house of Macmillan, which continued as Marshall's publisher thereafter. Ostensibly an elementary primer, this book contained the first general statement of Marshall's emerging theories, and a considerable sophistication lay beneath its deceptively simple surface. Together with the powerful *Pure Theory* chapters published by Sidgwick, a few copies of which circulated outside Cambridge, *The Economics of Industry* marked Marshall as a rising star in the economics firmament. With the death of W. S. Jevons in 1881, he moved into the public eye as the leader in Britain of the new scientific school of economics.

The duties of the Bristol principalship proved irksome to Marshall, especially as the college was struggling financially. He was anxious to proceed with his writing, having by 1877 conceived the plan for the book that was to become the *Principles*. His frustrations were increased by the onset in 1879 of a debilitating illness, diagnosed as kidney stones, which restricted his activities. He was persuaded to continue as principal until 1881, when he resigned both posts at the college. The next year was spent travelling, with an extended sojourn in Palermo, and it was in this year that composition of the new book began in earnest.

At Bristol, Marshall had got to know well Benjamin Jowett, the famed Master of Balliol, who was one of the governors of the struggling college. It was probably by Jowett's generosity that Marshall was able to return to Bristol in 1882 as Professor of Political Economy. And it was doubtless at Jowett's instigation that the Marshalls moved to Oxford in 1883, when a Balliol lectureship became vacant on the unexpected death of Arnold Toynbee. Marshall had considerable success as a teacher in Oxford and appeared settled in for an indefinite stay. But an 'Oxford School of Economics' was not to be. The sudden death of Henry Fawcett, who had been Professor of Political Economy at Cambridge since 1863, opened up the irresistible prospect of a return to Cambridge and a position with great potential for academic leadership. Marshall, the dominant candidate, was duly elected in December 1884, holding the chair until 1908, when he resigned to devote himself entirely to writing.

In many ways Cambridge's inviting prospects were to prove illusory. Economics was taught as part of the Historical and Moral Sciences Triposes, but neither avenue provided a supply of able interested students, nor was there much scope for advanced work. Marshall struggled for many years, with limited success, to increase the scope for economic teaching. But it was not until 1903, with the establishment of a new Tripos in Economics and Politics, that his goal was achieved. Even then, few resources were made available by the university and colleges for the teaching of economics, and the staffing of the new Tripos relied heavily on Marshall's willingness to support two young lecturers from his own pocket. The flowering of the new school came about mainly after his retirement, but the seeds were certainly planted by his efforts.

Absorbed in the struggle for his own subject, Marshall took relatively little part in general university affairs. Indeed, his rather obsessive

personality and proneness to magnify details would have made him ineffectual as a university statesman even if he had aspired in that direction. But he did play a prominent part in the successful campaign of 1896–7 against the granting of Cambridge degrees to students of the women's colleges – this despite his wife being at the time a lecturer at Newnham. He was not opposed to women's education, indeed had been a warm supporter in his early days, but was vehemently opposed to the assimilation of women into an educational system designed for men.

But the dominant fact in Marshall's life after his return to Cambridge, and certainly the aspect of greatest interest to posterity, is his long struggle to give adequate written expression to the stores of economic knowledge and understanding he had accumulated. The demands of teaching and administration left him little time or energy for sustained composition during term time and it was in the jealously guarded long vacations, usually spent away from Cambridge on the south coast of England or in the Tyrol of Austria, that the only real progress could be made. By 1887 the book commenced in 1881 had grown into a projected two-volume treatise. He hoped to complete the first volume in time for it to appear in the autumn of that year with the second volume appearing by 1889. In fact, the first volume (1890a) appeared as the *Principles of Economics, Volume One,* only in July 1890, when it was received with great and immediate acclaim and established Marshall firmly as one of the world's leading economists. The second volume never appeared. It was to have covered foreign trade, money, trade fluctuations, taxation, collectivism and aims for the future – a tall order!

Marshall struggled for the next 13 years with his intractable second volume, meanwhile spending much time on substantial, but not very substantive, recastings of the first volume in new editions of 1891, 1895 and 1898, and in preparing a digest of it to replace the earlier *Economics of Industry* which he had come to dislike intensely. (The digest, 1892, appeared under the title *Elements of the Economics of Industry, Volume One.* Like the earlier work it included material on trades unions that was never incorporated into *Principles.*) By 1903 much material had been accumulated for the second volume, but the scope was becoming unmanageable as Marshall became increasingly preoccupied with problems of trusts, trades unions, international trade, and comparative economic development, and decreasingly concerned with matters of pure theory. In that year, partly from the impetus of writing a private memorandum on trade policy for the use of the then Chancellor of the Exchequer, and partly because the tariff controversy was at full heat, Marshall was tempted into writing a short topical book on foreign trade questions, intending to publish it speedily. But this project too grew unmanageably in his hands. In 1907, the preface to the fifth edition of *Principles* (the last major rewriting) announced the abandonment of the proposed continuation and promised instead a volume, already partly in print, on 'National Industry and Trade', to be followed soon by a companion volume on 'Money, Credit and Employment' (Guillebaud 1961, vol. 2, p. 46). To reflect this change, the title of the sixth and subsequent editions of *Principles* was changed to *Principles of Economics: An Introductory Volume.* Retirement in 1908, at the age of 66, freed Marshall to concentrate on these projects, but progress continued to be slow. He appears to have suffered from recurrent dyspepsia and high blood pressure, necessitating a strict regimen and limiting his ability to work. But the more fundamental problem was that the world kept changing and the increasingly realistic and factual tone of his enquiry called for incessant recasting and revision. Nothing had been completed by the time war broke out in 1914, and then much rewriting was required to take into account the radical changes that were transforming the world economy and its post-war prospects. At last, when Marshall was 77 years old, *Industry and Trade* (1919), his second masterpiece, finally appeared. It was a magisterial, largely factual, consideration of trends in the British and international economy and of future economic prospects. But, lacking an obvious theoretical skeleton, it has not received from economists the kind of attention lavished on *Principles*, although interest in it is now beginning to stir among historians of economics.

In its final form, *Industry and Trade* was narrower in scope than had been intended earlier, while the proposed book on 'Money, Credit and Employment' still remained to be written. Over the next 4 years, by a remarkable effort, and despite rapidly waning powers, some of the mass of accumulated raw material remaining was pulled together in *Money, Credit and Commerce* (1923). This contains Marshall's fullest treatment of the theories of money and international trade, but it is an imperfect pastiche of earlier material, some dating back almost 50 years.

In the last months of his life, Marshall toyed with the occasional writings and the memoranda and evidence for governmental enquiries that he had prepared at various stages during his career, with the hope of editing them for publication in book form. This was not to be, but his plan was largely fulfilled after his death in two books sponsored by the Royal Economic Society (Pigou 1925; Keynes 1926).

Judged by what might have been, Marshall's authorial performance after 1890 was a sorry one, marked by repeated procrastination and inconstancy and by chronically over-optimistic expectations. The mantle of leadership that he had assumed on Jevons's death had proved a heavy one. Both temperamentally and by virtue of his acknowledged position as the doyen of British economists, Marshall was compelled to attempt the magisterial and to denigrate the kind of forceful direct essay of which he was eminently capable.

As Cambridge professor and unquestioned leader of British orthodox economists, Marshall could hardly avoid becoming a public figure whose pronouncements carried more than a personal weight. His consciousness of this, and of the precarious public standing of economics, as well as his own temperament, made him peculiarly reluctant to enter into public controversy, although he would on occasion fire off a letter to *The Times* on some issue of the day. He served as an expert witness for several government enquiries and was an influential member of the Royal Commission on Labour of 1890–94. As President of Section F of the British Association in 1890 he took the formal lead in the movement to found the British (later Royal) Economic Association, but he was not a prime mover. Indeed, he was not a clubbable or organizational man and relied on others to further whatever goals he desired for economics and the economics profession at large. But neither was he a recluse. Balliol Croft received a continuing stream of visitors, ranging from working class leaders to distinguished foreign economists, while students or young colleagues were always welcomed and offered generous advice mixed with exhortation.

Although able students interested in economics were in short supply, Marshall did over the years teach and influence several students who were to make contributions to the subject. From the early Cambridge period H.S. Foxwell, H.H. Cunynghame, J.N. Keynes and J.S. Nicholson might be mentioned. The Oxford period brought L.L.F.R. Price and E.C.K. Gonner, while the period as Professor in Cambridge produced, among others, A. Berry, A.W. Flux, C.P. Sanger, A.L. Bowley, S.J. Chapman, A.C. Pigou, J.H. Clapham, D.H. Macgregor, C.R. Fay, and, last but not least, J.M. Keynes.

The undoubted fact of Marshall's professional leadership of British economics calls for some explanation. He was far from suited to such a role by temperament, and his fussiness and inflexibility could be irritating. For example, Sidgwick, J.N. Keynes, and Foxwell, the most important of his early allies in Cambridge, were all eventually alienated. Marshall's success can be attributed partly to sheer persistence. As in the case of the new Tripos, he had a clear idea of what he wanted to accomplish and worried away at it until he exhausted the opposition and was allowed to have his way. But it must also have been due to the lack of any alternative. The relevant question is not 'Why Marshall?' but 'Who else?' Economics was rapidly evolving as a profession around the turn of the 20th century, creating a leadership vacuum. Leadership was unlikely to emanate from outside Oxford, Cambridge or London, but F.Y. Edgeworth at Oxford was perhaps the last man capable of meeting the need, while E. Cannan at the new London School of Economics, although more suited than Marshall to the hurly-burly of professional politics, was too

much the perennial critic and iconoclast to fill the bill. Moreover, whatever Marshall's foibles, the sheer power of his intellectual vision, his international standing as Britain's leading economic thinker, and his ability to inspire an impressive flow of budding scholars, all conspired to make him the only feasible contender.

Marshall's Views on the Social Setting, Aims and Methods of Economics

Marshall saw economics as concerned with those aspects of human behaviour open to pecuniary influences and sufficiently regular and ubiquitous to permit statements of broad scope and some persistence. While maintaining, especially in earlier work, that some heeded moral imperatives might be impervious to pecuniary considerations, he conceded that most behaviour lay within the ambit of the measuring rod of money. On the other hand, he emphasized that motivation was not merely a matter of pursuing pecuniary self-interest, even if broadly conceived to include interests of family and friends. He was anxious to lay the ghost of *homo economicus* and emphasized the human desires to obtain social approbation or distinction and to enjoy the pleasures of skilful activity. He saw actors as diverse as captains of industry and sculptors driven more by the joys of creative activity and the striving for the regard of peers than by the desire for material acquisition.

As well as not being pecuniary maximizers in any narrow sense, individuals were for the most part seen as imperfect optimizers. The working classes, especially, often lacked the knowledge and foresight to judge their long-term interests. Marshall's actors were not imbued with complete knowledge of their environment but had to acquire knowledge slowly, and often painfully, through experience. Nor were they endowed with fixed desires and an intrinsic, unchanging character. Indeed, character and preferences evolved as individuals were exposed to new possibilities and chose to enter into new activities. The workplace, in particular, was an important moulder of character. Self-improvement and character development induced by environmental changes, planned or unplanned, both figured largely in Marshall's world view. He believed that social institutions, such as land tenancy practices, were pliable and ultimately moulded themselves into conformity with the individual interests involved, rather than presenting a permanent constraint on mutually desired accommodations. (For this he was taken to task by his most vehement critic, W. Cunningham, who denied the applicability of modern economic theory to medieval practices – see Cunningham 1892.) But institutional change must be slow, slower even than changes in individual character and wants, because informal customs and tacit agreements are hard to change. Thus, while the institutions and informal understandings and prohibitions that constrain and mould economic behaviour might ultimately be endogenous they will often be ill adapted to current circumstances and thereby act as an independent constraint on the pursuit of mutually desired accommodations. Institutions, in the broad sense, are important and not always socially rational constraints on individual action.

Marshall was impelled to economics because 'the study of the causes of poverty is the study of the causes of the degradation of a large part of mankind' (1920, p. 3). For the bulk of the population, mired in poor living and working conditions, little progress in habits, aspirations and self-esteem could be expected without prior improvement in economic conditions. Such improvement was socially important not so much for its own sake, at least once the pangs of immediate want were assuaged, but because of its instrumental role in permitting and stimulating improvement in the quality and character of the population. What Marshall really valued was not improvement in the standard of living but the enhancement of the standard of life that this improvement made possible. And he entertained little doubt about what constituted a qualitative improvement here, even though – or perhaps because – his values may seem quite parochial and culture-bound.

Economic improvement required appropriate institutions, incentives and attitudes, and would be threatened by wide-scale government intrusions into economic affairs, although some forced

income redistribution could be tolerated. But even if economic conditions were improved, the full yield of social betterment would be garnered only if enlarged consumption were turned to ennobling and horizonexpanding channels (rather than, say, to strong drink), involved a due consumption of beneficial leisure, and was accompanied by healthier and less stultifying conditions of working and town life. The government had a guiding role to play here. But even more important would be the assistance and example of employers and the upper and middle classes, who must first rid themselves of a frequent propensity to showy and ostentatious consumption and excessive materialism. The working-class leaders and skilled artisans who had already raised their own standard of life had an important leadership role too. Voluntary individual efforts to assist the rise of the underprivileged must rest on an adequate understanding of economic consequences. For this, as well as to secure an informed electorate, the diffusion of sound economic knowledge was an essential and integral element in the process of socio-economic transformation. Economics thus was itself a noble activity of high importance for the future of mankind.

The broad view of the economy suggested by the foregoing is of a complex evolutionary process of combined economic, social and individual change in which each individual's abilities, character, preferences and knowledge develop jointly, along with social institutions, markets and the technologies of production and communication. The pursuit of self-interest, broadly conceived, is ubiquitous in directing this evolutionary process, but is subject to inertia, ignorance and limited foresight, not to mention individual mutability.

Unfortunately, Marshall was able to bring little formal analysis to bear on this general 'biological' vision of the economy and could only evoke it descriptively. It might be true that 'the Mecca of the economist lies in economic biology rather than in economic dynamics' (1920, p. xiv). Nevertheless, the only available analytical tools were those of classical mechanics, tools that Marshall's early mathematical training had equipped him to employ skillfully. In fact, chief reliance had to be put on that branch of classical mechanics dealing with statics. Dynamics, beyond a few qualitative applications, required more precise information than was likely to be available. Perforce then, much of Marshall's formal analysis, like that of W.S. Jevons or Léon Walras, was based on simple assumptions of individual optimization and market equilibrium, taking preferences, technology and market institutions for granted. Such provisional or tentative 'statical' treatments could often be valuable. Indeed Marshall viewed them as indispensable for the correct analysis of many questions. But he was always anxious to stress that the analysis was preliminary, and perhaps of only transitory validity. This awareness made him impatient of overelaboration, so that, for example, he showed no interest at all in pushing the statical approach to its logical conclusion in the general equilibrium analysis of the stationary state. For him, equilibrium analysis was an indispensable but rough and ready instrument that needed to be employed with due caution and a continuing awareness of its limitations in the face of a complex ever-evolving reality. It was only a tool and did not itself constitute concrete knowledge.

Marshall had no great profundity as a philosopher of science and had little patience with metaphysics: 'in a sense … he held no views on method' (Coase 1975, p. 27). Marshall's discussions of methodology largely reflect the philosophical presuppositions of his day. His method was in the general deductive tradition of John Stuart Mill, but he sought to emphasize the relativity of particular theories, as contrasted with the universality of the general theoretical 'organon' or economists' toolbox. Anxious to present a public image of the unity of economics in the face of the Methodenstreit among economists in the late 19th century, he attempted to maintain an uneasy balance on method, decrying extended chains of deductive reasoning but denying the possibility of purely inductive inference unguided by a coherent conceptual framework. Economics had room for specialists in both deductive and inductive methods, but both must ultimately be co-workers. Assumptions must be selected with close regard to the facts of the case and potential disturbing causes must be kept prominently in mind and due allowance made for them.

J.N. Keynes described Marshall's analytical method as 'deductive political economy guided by observation' (1891, p. 217n) and Keynes's chapter 'On the Deductive Method in Political Economy' (1891, pp. 204–35) is perhaps as good a rationalization of Marshall's method as one can find.

Intellectual Debts

The intellectual background to Marshall's work in economics was established in the 1860s, partly in his stringent mathematical training, but perhaps more importantly in the heady mixture of utilitarianism, evolutionism and German idealism which he eagerly imbibed in the years immediately following his graduation. He seems to have started on economics from J.S. Mill's *Principles of Political Economy* (1848), moving on to the classic works of Smith and Ricardo. At a fairly early stage, probably around 1868, he discovered Cournot's *Récherches* (1838), which provided examples of the application of mathematics to economic questions. Acquaintance with J.H. von Thünen's work, which influenced Marshall's distribution theory, must have come somewhat later, in the early to mid-1870s. During the 1870s and early 1880s Marshall also read widely on economic development and socialism, including much literature in German, the only foreign language he mastered thoroughly. After that, his reading seems to have been concentrated mainly on factual and practical matters. Once his own theoretical views had crystallized, he appears to have been reluctant to do more than attempt to explain and clarify them to others, and to have taken remarkably little interest in new theoretical issues or in the theoretical ideas of others.

In many ways, the list of Marshall's denials of theoretical indebtedness is more remarkable than that of his acknowledgments. He claimed to have developed his ideas on consumer surplus before learning of anticipations by J. Dupuit and H. Fleeming Jenkin. The grudging attitude to W.S. Jevons's marginal utility theory shown in his review (1872) of Jevons (1871), although subsequently relaxed, was never replaced by any acknowledgement of indebtedness. He showed little or no interest in the work of Walras, gave meagre credit to Carl Menger, whose work must have become known to him by the early 1880s, patronized Pantaleoni and Böhm- Bawerk, largely ignored Pareto, and so on. Even in the case of Edgeworth, one of his few intimates, Marshall felt that undoubted theoretical powers were guided by an unreliable judgement and refused to follow Edgeworth's subtle elaborations far. In fact, the only major theorist of the day to command Marshall's entire admiration and respect was J.B. Clark, and even here there was no acknowledgement of serious indebtedness. This tendency to denigrate the work of his contemporaries was matched by an equally strong tendency to overvalue the achievements of the British Classical School led by A. Smith, D. Ricardo and J.S. Mill. For one reason or another – perhaps a personality quirk, perhaps an effort to boost the public esteem of economics – Marshall was prone to exaggerate the intellectual continuity and maturity of his subject – see O'Brien (1990) on this.

A growing interest in wider intellectual influences on Marshall in his formative years 1865–70 has been sparked by the publication and analysis of his early philosophical manuscripts (Raffaelli 1994, 2003), especially a paper entitled 'Ye Machine' that outlines a mechanism capable of learning new routines from experience, thus freeing its limited learning ability to gradually establish new and higher level routines, and so on. It appears that Marshall's ambitions in these early years lay in the area of 'psychology' or perhaps better in the 'philosophy of mind'. Whether the world lost more than economics gained from his switch to economics remains an open and perhaps insoluble question. But it does appear that the pattern of a sequential routinizing of new methods, continually leading to new levels of individual or organizational complexity, continued to play a significant part in Marshall's economic thought. More generally it is clear that he read philosophical literature widely in his formative years: Kant, Hegel, H. Spencer, and others. But whether and how these sources influenced his economic thought remains uncertain, partly because evidence is slight or absent.

Demand Theory

So far the discussion has remained on a very general level, dealing with broad aspects of Marshall's life and work. At this point there begins a much more detailed and technical consideration of various aspects of his theoretical contributions, commencing with his demand theory. Marshall's treatment of the theory of demand is sketchy and incomplete, concentrating on the demand for a single commodity, or commodity group, against a loosely defined background. A utility-maximizing individual's utility is defined by $u(x) + w(y)$ where x is the individual's consumption of the particular good X, while y is the individual's expenditure on all other goods. This expenditure is measured in money of constant purchasing power: that is, deflated by a general price index. How this index is defined and whether, as seems appropriate, the price of X is excluded from it, is left unclear. Such money can be treated as a composite good, Y, and y can be regarded as the amount of this composite good consumed. If m is the individual's initial endowment of Y, then $y = m$ whenever $x = 0$, while if X can be freely purchased at a fixed price of p units of Y per unit of X then x and y must satisfy the constraint $px + y = m$. Marshall assumes that the utility functions $u(x)$ and $w(y)$ have positive but diminishing marginal utility so that $u'(x) > 0 > u''(x)$ and $w'(y) > 0 > w''(y)$, where single and double primes are used to denote first and second derivatives. The maximum expenditure, e, that the individual is willing to make to secure x units of X is implicitly defined as a function e(x, m) by $u(x) + w(m - e) - w(m) = 0$. Providing that x and y are both positive, the rate at which e increases with x is $u'(x)/w'(m - e)$ by the implicit function theorem. This ratio would be the demand price for the xth unit of X if all previous units had been acquired at their corresponding demand prices: that is, if the individual had faced perfect price discrimination in exchanging Y for X. Alternatively, if the individual had been able to obtain any amount of X at fixed per unit price, p, the resulting demand function $x(p, m)$ for X would be implicitly defined (given x and y are both positive) by the first-order condition $u'(x) - pw'(m - px) = 0$. Partial differentiation of $x(p, m)$ shows that x falls as p increases, while an increase in m increases both x and y: thus, the Giffen possibility of an increase in p increasing the quantity of X demanded is excluded. But an increase in p may lower or raise the value of px, so that demand for X may be price elastic or price inelastic at a given p.

The possibility of buying at a fixed price rather than facing perfect price discrimination creates a consumer surplus of $e(x(p, m), m) - px(p, \text{m})$. This is the additional amount that could have been extracted by perfect price discrimination for all units up to the price-taking optimal one. That this surplus is positive follows from the fact that every infra-marginal unit of X acquired creates a surplus utility when the individual faces a fixed price (since $u'(x) > pw'(m - px)$ for each such x) but no surplus when the individual is faced with perfect price discrimination.

Marshall's mathematical notes (1920, pp. 838–42) on his general case are obscure and puzzling. Doubtless he felt this case was too dependent on unobservables to be of much practical value. He therefore emphasized the special case in which the marginal utility of money is treated as a constant. The rationale offered is that an individual's 'expenditure on any one thing ... is only a small part of his whole expenditure' (1920, p. 842). This simplifies $e = e(x, m)$ above to $e = u(x)/w'(m)$ while $x(p, m)$ is now defined implicitly by $u'(x)/w'(m) = \text{p}$. At the x value defined by the latter equation, consumer surplus arising from the ability to buy any amount of X at the per-unit price p can be expressed in utility terms as $u(x) - xu'(x)$ or in money terms as $u(x)/w'(m) - xu'(x)/W(m)$. These formulae are exactly analogous to the standard formula for Ricardian land rent, with the first term the output obtained on a piece of land from the application of x doses of variable input, each dose remunerated at the common marginal product. Partly because of this analogy, Marshall used the term 'consumer rent' rather than 'consumer surplus' prior to 1898.

Although priority must go to Dupuit, Marshall's simple concept of consumer surplus based on the assumption of a constant marginal

utility of money has been influential. But he was well aware of the complications arising from variation in the marginal utility of money: 'Strictly speaking we ought to take account of the fact that if he spent less on tea the marginal utility of money to him would be less than it is, and he would get an element of consumers' surplus from buying other things at prices which now yield him no such rent' (1920, p. 842). Although such influences may be 'of the second order of smallness' they raise the more disquieting issue of assessing the overall welfare effects of changes that affect many markets simultaneously. On this Marshall had little to say: 'the task of adding together the total utilities of all commodities, so as to obtain the aggregate of the total utility of all wealth, is beyond the range of any but the most elaborate mathematical formulae' (1920, p. 131n.). It was a task he chose not to pursue. Apart from generalizing for the possibility that a certain quantity of good X might be indispensable, Marshall elected not to develop his demand theory further, or even to generalize it to incorporate utility functions that were not additively separable (1920, p. 845). It is clear that each commodity in turn might take the spotlighted role of good X and that in certain circumstances simultaneous consumer surpluses for several goods might be added (1920, p. 842). An unpublished early manuscript note from the 1870s on the theory of taxation (Whitaker 1975, vol. 2, pp. 285–305) had advanced matters considerably further by working formally with the maximization of utility under a budget constraint, but this lead was not followed up in print and some of its lessons for welfare economics were apparently forgotten. *Principles* gave a clear intuitive account of the consumer's overall optimization problem (1920, pp. 117–23), but failed to connect it to the resulting interrelated set of demand functions for the various goods consumed. Indeed, it is clear that for positive purposes Marshall was willing to treat market demand functions in a quite pragmatic way, admitting, for example, close substitutes or complements and the Giffen exception, all inconsistent with the simple formal theory set out above. In judging this, it must be borne in mind that consistency and generality of 'statical' analysis were not Marshall's real goal. Rather, 'fragmentary statical hypotheses are used as temporary auxiliaries to dynamical – or rather biological – conceptions' (1920, p. xv).

The market demand for a good that is offered to all actual or potential buyers at the same given price is of course obtained as a function of that price by summing the amounts demanded at that price by all the consumers. A sufficient but not necessary condition for market demand to fall as price increases is that each individual's demand decreases. The now familiar concept of market demand elasticity – proportional quantity change divided by proportional price change – was first introduced by Marshall, although several authors had come close to the idea previously. It appeared without flourish in (1885c), and appeared more prominently in *Principles.* But Marshall himself made relatively little use of it.

Bibliographic note: Marshall's treatment of demand is essentially contained in (1920, pp. 92–137, 838–43). An influential, although controversial, interpretation of Marshall's demand theory is given by Friedman (1949). Biswas (1977) gives another alternative to the orthodox reading provided by Stigler (1950) that is largely adopted here. An excellent overview is Aldrich (1996). On consumer surplus see Chipman (1990).

Production and Long-Period Competitive Supply

In deriving the long-period supply curve of a commodity in *Principles,* Marshall envisages production as organized by firms, typically family businesses. Each firm strives to minimize its production costs, substituting one productive factor or production method for another according to the 'Principle of Substitution'. In its simpler forms this involves marginalist adjustment to bring relative marginal value products into line with relative marginal costs. But more generally, the Principle of Substitution is akin to a natural selection process, being 'a special and limited application of the law of survival of the fittest'

(1920, p. 597). Marshall's firms do not have costless access to a common production function, but must grope and experiment their way to cost-reducing modifications. The long-period supply curve is defined for a given state of general scientific and technical knowledge. But each firm must explore this to some extent anew.

Although the distinction is not entirely clear – distinctions seldom are for Marshall – two polar cases may be distinguished within his theory of long-period competitive supply. These will be referred to as the 'agricultural' and the 'industrial' cases. The former is much the more straightforward and involves an industry in which production is relatively simple, internal economies of scale are minimal, and the product is homogeneous and easily marketed. The optimal firm size is small, and management is sufficiently routine to need no exceptional ability to keep a firm operating efficiently. As the overall market expands, new firms may be added, but changing composition of the population of firms is not an essential feature of this case.

The long-period supply price per unit of output at which such an industry can supply any quantity of output must just cover the cost of maintaining that level of output indefinitely. That is, it must just suffice to pay all the inputs (including management) needed to produce that level of output in a cost-minimizing way at rates that just ensure that the requisite input quantities will continue to be forthcoming indefinitely. In the case of skilled workers, in particular, the rate must just suffice to induce parents to apprentice new workers to the industry at a rate exactly offsetting the attrition through retirement and other causes. Similarly, the return to fixed capital must just suffice to induce replacement of the existing stock of fixed assets, while the return to management must keep up the necessary replacement flow of managers. On the other hand, the return to land services must just suffice to prevent these services from migrating elsewhere, replacement not being necessary. As the level of industry output being considered is increased, the supply price will probably rise, mainly because of the need to pay a higher return to land so as to attract a greater supply from other uses, but perhaps also because of the need to pay more for rare natural talents that, like land, must be attracted in greater quantity from other uses, not being capable of replication through education and training. Such a tendency for long-period supply price to rise with output may be mitigated though seldom eliminated by substitution against inputs whose supply price is rising, and by possible external economies that increase each firm's efficiency by influences that depend, not on its own output, but on the entire industry's output. A tendency for supply price to rise with output will imply that infra-marginal units of those inputs whose supply prices are rising receive rents, since all units will be remunerated at the rate necessary to induce continuing supply of the marginal unit. In the absence of external economies (or diseconomies) the total rent or producer surplus generated will be the 'triangular' area above the supply curve. That is, it will be

$$R = xg(x) - \int_0^x g(v)\mathrm{d}v$$

where $g(x)$ is the supply price of output quantity x, an increasing function of x. This result does not apply in the presence of external economies. In later editions of the *Principles,* Marshall introduced the device of the 'particular expenses curve' (1920, pp. 810–12) to display rent in such a case, but this *ex post* construction does not give an independent basis for determining rent.

It that the long-period supply curve of an industry depends on the general economic background against which the industry is assumed to operate. As is the case with demand, Marshall does not consider this background in detail. He assumes prices to be expressed in money of constant purchasing power and recognizes on occasion that there may be close interrelations between two industries (for example, they may compete for the same specialized land). He also recognizes that 'a theoretically perfect long period must give time enough to enable not only the factors of production of the commodity to be adjusted to the demand, but also for the factors of production of those factors of production to be adjusted and so on' (1920, p. 379n.), and that this leads ultimately to the assumption of a stationary state. But

he is not willing to follow this route far and is content in general to take the supply conditions of the factors of production for granted when analyzing long-period price determination.

In the 'manufacturing' case, to which we now turn, the product is differentiated, marketing is difficult, and each firm must build up and retain goodwill and a customer connection for its own specialized product. There are substantial internal economies of scale in production and successful management calls for business ability of a high and rare character. In this environment, a family business may be built up by an exceptional founder, but this build-up must be slow because of the difficulty of establishing a market and perhaps also because of constraints on financing. And when the founder passes on, his successors are unlikely to have equal talents or even the lesser talents required to prevent the firm's business from languishing. By the third generation of succession, the firm is likely to expire. Even a joint stock company (a case added rather as an afterthought) is likely to ossify into bureaucratic stagnation, and presumably the same is true of family businesses that rely on paid managers. Thus, the typical firm in the manufacturing case passes through a finite life cycle, and the industry is comprised of a population of such firms at various phases of the life cycle, some in the early expanding phase, others in decline.

The long-period supply price at which such an industry can supply a specified level of output must now be regarded as an index of the prices of all the different firms' products. It must meet all the conditions required in the agricultural case. Thus, the price must allow for a continuing replacement flow of the various types of workers (including managers) and fixed assets, as well as the retention of the necessary 'land' services. But now there must also be a surplus sufficient to induce are placement flow of new firms – a supply of 'business organization' that will just suffice to replace the expiring firms and keep the age distribution of firms constant.

Industry equilibrium does not require each firm to be in an unchanging equilibrium any more than the trees in the proverbial forest. A new firm will be established if the prospective earnings over the expected life cycle appear to justify the cost and trouble involved. The firm's initial earnings are likely to be negative as it slowly builds up its technical expertise and market connections, but these early losses can be regarded as investments to be recouped in the later stages of the firm's prospective life cycle.

It is here that Marshall's 'representative firm' enters the picture. It is best regarded as a parable that avoids the need to consider the entire distribution of firms. By definition, the long-period supply price of any level of industry output is the average cost of the representative firm at that level of output. Industry-level magnitudes may then be regarded as if they were generated by a fixed number of unchanging representative firms rather than by the actual heterogeneous body of ever-changing firms. In other words, the manufacturing case may be treated as if it were an agricultural case populated by representative firms only. Such arguments add nothing conceptually and are prone to confuse, although it might be noted that Marshall believed an acute well-informed observer could select an actual firm that was close to being representative in this sense.

The average cost and size of the representative firm will change as industry output changes. There are two main reasons for this. A larger industry output is likely to generate more external economies, lowering the costs of every firm. But more importantly, the larger industry demand is, the easier it will be for a new firm to build up a market, and so the larger the size to which firms will grow before they begin to decline. This will bring about greater access on average to unexhausted internal economies of scale, again leading to lower costs on average. For both these reasons, long-period supply price is likely to decline as a larger industry output is considered, even though the opportunity cost of obtaining greater supplies of land services and rare natural talents may rise. Again, the particular expenses curve may be used to display the producer surpluses or rents accruing to such scarce factors at any given level of industry output, but the relationship of this family of curves to the long-period supply curve is tenuous and complex. Rent obviously cannot be represented by a 'welfare triangle' above the supply curve when the latter is falling.

The conception of competition in Marshall's manufacturing case is much closer to later ideas of imperfect or monopolistic competition than to modern notions of perfect competition. Products are differentiated and firms are not price takers, but face at any time downward-sloping demand curves in their special markets. Even if the difficulties of rapidly building up a firm's internal organization can be overcome, the resulting enlarged output cannot be sold at a price covering cost – even granted substantial scale economies in production – without going through the slow process of building up a clientele and shifting the firm's particular demand curve. The time this takes is assumed to be considerable relative to the duration of the firm's initial vitality. But in some cases the difficulties of rapid expansion may be overcome. They may not have been very severe, as when different firms' products are highly substitutable, or the firm's founder may have unusual genius. In such cases the industry will pass into a monopoly or be dominated by a few, strategically interacting firms, or 'conditional monopolies' as Marshall termed them.

Marshall's reconciliation of persisting competition with increasing returns and falling supply price is complex and problematic, but it does not depend in any essential way on scale economies being external to the firm. The concept of external economies is one of his significant contributions, although his treatment of it can hardly be called pellucid. But it was added more for verisimilitude than because it was theoretically essential to the structure of his theory.

The issues surrounding Marshall's representative firm, and the problem of reconciling the persistence of competition with the presence of unexhausted internal economies of scale, continue to receive attention among historians of economic thought but no definitive reading has yet been attained, or perhaps ever will be. The account given above is well supported by Marshall's text, but as is often the case with Marshall, elements of ambiguity and vagueness remain.

Bibliographic note: Marshall's treatment of long-period competitive supply is to be found in (1920, pp. 314–22, 337–80, 455–61, 805–12) and (1919, pp. 178–96). The earliest version, dating from the early 1870s is reproduced in Whitaker (1975, vol. 1, pp. 119–59) and see also (1879a). Key early commentaries and criticisms of Marshall's theory of supply are Sraffa (1926), Robbins (1928), Robertson et al. (1930), Viner (1931), Frisch (1950), Hague (1958) and Newman (1960).

Price Determination and Period Analysis

The long-period supply curve for any good indicates for each market quantity the least price at which that quantity will continue to be supplied indefinitely. The long-period equilibrium price and quantity are determined by the intersection of this supply curve with the market demand curve, assumed to be negatively sloped, that indicates the highest uniform price at which any total quantity can be sold. In the agricultural case, equilibrium will be unique as the supply curve slopes positively. But in the manufacturing case, the supply curve, as well as the demand curve, may have negative slope, so that multiple equilibria can occur. Equilibrium is adjudged locally stable if demand price is above (below) supply price at a quantity just below (above) the equilibrium quantity. The intuitive justification for this is that the actual price of any available quantity is determined by the demand price, while quantity produced tends to increase (through both expansion of existing firms and entry of new firms) whenever an excess of market price over supply price promises high profits, while it tends to decrease in the opposite case.

This stability argument is sketchy and, in any case, there still remains the question of exactly how a new long-period equilibrium is attained following some change, such as a permanent shift in the demand curve. One possibility would be to consider explicitly the adjustment process through time, but Marshall preferred to approach the problem by another route – his period analysis, one of his most memorable and lasting contributions. (His passing claim (1920, p. 808) that the long- period supply curve may not be reversible, supply price depending upon past-peak output as well as current output, is something of an

exception to this generalization. It appears to rest on some restriction of the degree of downward supply adjustment, and so not to involve a true long-period analysis, or else to invoke a kind of learning by doing that once attained is not readily lost.)

Period analysis is Marshall's most explicit and self-conscious application of the comparative-static, partial-equilibrium method with which his name will always be associated. As he observed,

> the most important among the many uses of this method is to classify forces with reference to the time which they require for their work; and to impound in *Caeteris Paribus* those forces which are of minor importance relatively to the particular time we have in view. (Guillebaud 1961, vol. 2, p. 67)

Which forces or variables are to be hypothetically frozen or impounded, and which are to be determined by the requirements of equilibrium (an equilibrium contingent upon the contents of the *ceteris paribus* pound, of course), should be determined pragmatically in each case with the aim of focusing on the features deemed dominant in that case. As a general rule, those forces should be impounded which move very slowly, or else bounce around very rapidly, relative to the length of 'the particular time we have in view'. This is well illustrated by Marshall's example of a fish market, where the focus may be on the determinants of price over a few days, a few months, or several years, or even decades (1920, pp. 369–71). As an expositional matter, however, and also to embody distinctions of wide (but not universal) applicability, Marshall emphasized three broad cases. Temporary or market equilibrium analysis proceeded on the assumption of a fixed stock of output already available or in the pipeline. Short-period normal equilibrium analysis permitted output to be varied, but not the stock of productive 'appliances' available to produce that output. 'Appliances' must be taken here to cover skilled labour and business organization as well as fixed capital assets, so that the existing set of firms is to be taken as given. Finally, long-period normal equilibrium, which has already been considered, allows the stock of appliances, as well as the level of output, to be freely varied. In this case equilibrium incorporates the conditions necessary for inducing an exact replacement flow of each kind of appliance, including a replacement flow of new firms in the manufacturing case.

Temporary equilibrium for a perishable commodity is simply a matter of selling off the existing stock. Marshall recognizes the possibility of 'false trading' – sales at a non-equilibrium price – but argues that (*a*) this will not affect the eventual price if the marginal utility of money is constant, and (*b*) price will quickly settle close to the uniform price that would just clear the market if used in all transactions. With a storable good there is the further speculative possibility of holding back supply for future sale, and this gives expected future cost of production an indirect role in influencing current market price. Cost of production already incurred is an irrelevant bygone, however.

In short-period normal equilibrium, output is adapted to demand within the constraints set by the fixed supply of available 'appliances'. High demand will raise equilibrium output, but only within the limits possible by working existing appliances more intensively or pulling in versatile unspecialized labour and land from elsewhere. Low demand will lead to low utilization of appliances, perhaps idleness of some, and migration of unspecialized inputs to elsewhere. In the agricultural case a firm will change output until marginal prime or variable cost equals market price. In the manufacturing case, a fear of spoiling the future market or invoking retaliation from competitors tends to make a firm's output more responsive to variation in market price, and hence to make market price less responsive to demand shifts. Otherwise, the two cases are similar, both involving a fixed population of firms and a rising supply curve.

The return received by an appliance will often exceed the minimum necessary to induce its operation at the chosen intensity (its prime cost) and this excess is a 'quasi rent'. To the extent that land and rare natural talents are immobile in the short period, or less mobile in the short period than the long, their returns too will often have a quasi-rent element. Otherwise, they will receive only

differential rents, though often at rates differing from their long-period values. It should be stressed that the concepts of quasi-rent and differential rent are relative to a specific use. The prime cost necessary to retain an input in this use may itself include rent or quasi-rent when viewed in the context of a more inclusive set of alternative uses. Thus, from the viewpoint of all possible uses in the economy, the return to any factor in fixed supply is entirely a rent or quasi-rent (the latter if fixity is only short-period).

Marshall paid little attention to the possibility that forces similar to those constraining the adjustment of supply when time is limited might also operate on the side of demand. Thus the same considerations underlie the market demand curve whether it is coupled with a temporary, short-period or long-period supply curve. In each case, market equilibrium price and quantity are determined by the intersection of the appropriate demand and supply curve. The stability of temporary equilibrium is directly asserted. The stability of short-period equilibrium depends on the same quantity-adjustment argument invoked for long-period equilibrium, but since the short-period supply curve is always positively sloped, uniqueness and stability are assured.

The theory of short-period normal equilibrium was designed as a tool for analysing unemployment and economic fluctuations in the never-completed second volume of *Principles*. But it also has use in explaining adjustment to a permanent disturbance. Suppose, for instance, that an industry is in long-period equilibrium when a permanent shift in demand occurs. The immediate or short-period effects can be analysed by freezing output and stocks of appliances at their initial levels. Insight into the actual adjustment through time can then be obtained by appropriately changing the output level assumed in the temporary equilibrium, so that movement of temporary equilibrium towards short-period equilibrium can be traced out as output, but not stocks of appliances, adjusts. Similarly, the levels assumed for the stocks of appliances in this short-period equilibrium can be allowed to change and the movement of short-period equilibrium towards long-period equilibrium traced out. Such arguments are now a staple of elementary pedagogy. They clearly require additional assumptions about the adjustment of output and the way in which investment or disinvestment in appliances proceeds, and are only a poor and ambiguous substitute for an explicit dynamic analysis. But such 'statical' procedures, although imperfect, may, in Marshall's words, be 'the first step towards a provisional and partial solution in problems so complex that a complete dynamical solution is beyond our attainment' (Pigou 1925, p. 312).

Marshall's period analysis, and more generally his partial-equilibrium approach to price determination, was designed in large part as a usable tool for the analysis of concrete issues. Its longevity amply testifies to its usefulness in this respect. But it was also meant to serve the more doctrinal purpose of clarifying the respective roles utility and cost of production play in determining value. The aim was to show that the greater the scope for supply adjustment permitted in the definition of equilibrium, the more dominant the supply side influence on price becomes. This doctrinal goal helps to account for the rather heavy weight given to long-period analysis in *Principles*. For, as Marshall recognized, its value as a tool of applied analysis is seriously qualified by the fact that 'violence is required for keeping broad forces in the pound of Caeteris Paribus during, say, a whole generation, on the ground that they have only an indirect bearing on the question in hand' (1920, p. 379n). That is, there is no good ground for assuming that background forces such as technology and tastes will remain constant for the length of time required for long-period equilibrium to be practically relevant. For concrete analysis of problems of such long duration it will often be necessary to transcend the period analysis, with its reliance on statical equilibrium, and undertake directly an analysis of secular change, of which Book 6, Ch. 12 of *Principles* on the 'General Influence of Economic Progress' (1920, pp. 668–88) offers the main example, but not a very impressive one.

In emphasizing the role that cost of production plays in the determination of long-period value, Marshall was not content to rest on money costs of production but sought to go behind these costs to

the real costs –the efforts and abstinences – for which in a non-coercive economy the money costs are recompense. In doing so he purported to follow Ricardian tradition, but is more plausibly viewed as attempting to place the newer subjective value theories in broader (but still subjective) focus. Just as the price paid by a consumer serves as a measure of marginal utility, with a consumer surplus gained on infra-marginal units, so the unit price received by a worker or saver measures the real cost or disutility at the margin, with a producer surplus on the inframarginal units of effort or abstinence. But, as Marshall recognized, the parallel holds imperfectly in the long period when workers must be regarded as produced means of production as well as final consumers and cost bearers. In particular, parental sacrifice for raising and training offspring obtains little or no direct pecuniary reward.

Bibliographic note: Marshall's treatment of period analysis is concentrated in (1920, pp. 363–80) but see Whitaker (1975, vol. I, pp. 119–59) for the earliest version. For commentary and exposition see especially Viner (1931), Opie (1931), Frisch (1950) and Whitaker (1982). On temporary equilibrium see (1920, pp. 331–6, 791–3, 844–5) and Walker (1969). On short-period normal value see Gee (1983).

Normal Value and Normal Profit

Implicit in the preceding discussion are Marshall's conceptions of normal value and normal profit. Normal value is defined as the value that would result 'if the economic conditions under view had time to work out undisturbed their full effect' (1920, p. vii). It is contrasted with market value, which is 'the actual value at any time' (1920, p. 349). Normal value is hypothetical, resting on a *ceteris paribus* condition, its role being to indicate underlying tendencies. The normal value of a commodity may approximate its average value over periods sufficiently long for the 'fitful and irregular causes' (1920, pp. 349–50) that dominate market value to cancel out, but this should not be presupposed automatically outside a hypothetical stationary state.

The distinction between normal and market value is closely related to the distinction between natural and market value found in the work of Smith and the classical economists. In 1879 Marshall had identified normal value with 'the results which competition would bring about in the long run' (1879b, p. vii), but in *Principles* he switched to the view that 'Normal does not mean Competitive' (1920, p. 347) and admitted any kind of regular influence so long as it was sufficiently persistent. The economic forces hypothetically permitted to achieve full mutual accommodation could now be chosen appropriately for each case. In particular, the distinction between short-period and long-period normal (or 'sub-normal' and 'true- normal' in earlier editions) was emphasized.

Profit was viewed by Marshall as the residual income accruing to a firm's owner, a return on the investment of the owner's own capital and recompense for the pains of exercising 'business power' in planning, supervision and control. Normal profit is essentially an opportunity cost, the minimum return necessary to secure the owner's inputs to their current use, or rather to accomplish this for an owner of normal ability. Marshall presumes that there is a large and elastic supply of versatile actual or potential owner managers of normal ability. In long-period equilibrium each of these must just receive the same normal rates of return on investment and exercise of business power whatever the line of business. However, those who are exceptional may do better, essentially by exerting greater business power.

These common rates of normal return are simultaneously determined, along with the normal returns to other kinds of effort and abstinence, by Marshall's macroeconomic theory of the long-period determination of factor incomes (see below). Although it is the case that profits are a residual, rather than a contractually agreed amount like other incomes, this difference is immaterial in long-period equilibrium. In particular, a long-period equilibrium analogy between ordinary wages and the normal earnings of business power is stressed. Normal profit is a necessary element in the costs that underlie the long-period normal supply curve, but actual profit is a quasi-rent or producer surplus for shorter periods.

Normal profits are a return to 'business power in command of capital' and compensate for three distinct elements: 'the supply of capital, the supply of the business power to manage it, and the supply of the organization by which the two are brought together and made effective for production' (1920, p. 596). The combined compensation of the latter two components comprises 'gross earnings of management', the return to the second component being 'net earnings of management'. In long-period equilibrium, the normal return to the first element is imputed at the market interest rate on default-free loans, and that to the second component at the rate paid to hired managers performing comparable tasks. The residual third element, the return to 'organization', is most straightforwardly interpreted as an extra return on owned capital equivalent to the premium for default risk, or 'personal risk', that would have to be paid on borrowed capital. In the manufacturing case, the annual level of normal profit for each firm in an industry must be interpreted as the annualized equivalent of the expected stream of returns that is just sufficient to induce an individual of normal ability to found a firm in the industry rather than divert energies and capital elsewhere. Normal ability here is defined relative to other potential founders of firms, a group already exceptional relative to the population as a whole. By construction, such normal profits must be earned by the representative firm.

Bibliographic note: The most pertinent commentary is Frisch (1950). For Marshall's views on normal value see (1879b, pp. v–vii, 65–71, 146–9; 1920, pp. vii, 33–6, 337–50, 363–80). For his views on normal profit see (1879b, pp. 135–45; 1920, pp. 73–4, 291–313, 596–628). For the role of 'personal risk' see Guillebaud (1961, vol. 2, p. 672).

Welfare Economics

To serve as a tool of welfare economics, monetary measures of consumer surplus, producer surplus and rent must be aggregated over individuals. But how are the resulting sums to be interpreted? Marshall's very limited and proximate attempts at formal welfare arguments are carried out within a utilitarian framework, for which the goal is maximizing aggregate utility. He implies that interpersonal utility comparisons are possible in principle and that utility functions will be similar for all members of any group that is homogeneous in terms of mental, physical and social attributes. Within such a group, the marginal utility of money will be the same for two individuals having the same income, and lower for the richer of two individuals with differing incomes, on the assumption in each case that both individuals face the same trading opportunities. A postulated government action may impose gains and losses on various individuals that can be measured and aggregated in money- equivalent terms. But how can these measures be translated into statements about aggregate gains and losses of utility? Marshall emphasizes two special cases. First, if the gains and losses are both proportionately distributed over income classes in exactly the same way, then net aggregate gain (positive or negative) in money will serve as an ordinal index for the net aggregate gain in utility. A corollary of this is that if two alternative actions affecting the same group have the same relative distributions of gains and losses over income classes then the alternative yielding the greater net aggregate gain in money must have the greater net aggregate gain in utility. Second, if some change makes for a zero net aggregate change in money terms, but the gains accrue to individuals of lower income than those bearing the costs, then the aggregate net utility gain must be positive – a warrant for certain redistributive policies. In other cases he sees that careful assessments of the marginal utility of money to the various injured and benefited groups would be needed, assessments that could be used to transform monetary gains and losses into utility measures. He toys (1920, pp. 135, 842–3) with using the Bernoulli hypothesis on the relation between wealth and utility as a basis for such calculations, but gives little indication as to how assessments might be made in practice.

Marshall's best known and most successful foray into formal welfare analysis was his proof that total welfare might be increased by using the proceeds of a tax on an 'agricultural' industry to subsidize a 'manufacturing' industry.

All comparisons involved long-period equilibria and relied on the validity of aggregated money-equivalent measures of gains and losses. He demonstrated that the gain in consumer surplus in the expanded decreasing-cost manufacturing industry might exceed the combined loss in consumer and producer surplus in the contracted increasing-cost agricultural industry. No formal account was taken of a possible gain in producer surplus in the manufacturing industry as this merely makes the argument hold *a fortiori.* The crucial point in this argument, as Marshall recognized, is that producers are not harmed by 'a fall in price which results from improvements in industrial organization' (1920, p. 472). It is immaterial whether the improved organization of the enlarged manufacturing industry is due to external economies or to internal economies resulting from an increase in the size of the representative firm. Contrary to much subsequent opinion, Marshall's tax-subsidy argument is not necessarily dependent upon external economies.

Another significant, but overlooked, welfare analysis provided by Marshall was that of a monopolistic public enterprise in a situation where taxation involves an excess burden (1920, pp. 487–93, 857–8). In the absence of this excess burden Marshall proposes that the enterprise seek to maximize 'total benefit', the sum of net profit and consumer surplus. This implies marginal cost pricing, since the area below the demand curve and above the marginal cost curve is maximized when the two curves intersect. But, given that taxation involves an excess burden, it may be desirable to augment tax receipts from monopoly revenue if the sacrifice of consumer surplus is small. Marshall proposes the alternative goal of maximizing 'compromise benefit', the sum of consumer surplus and monopoly revenue when the latter is in effect multiplied by the marginal cost of raising a unit of government revenue from other sources. Maximization of compromise benefit leads to the setting of what has come to be termed a 'Ramsey price'.

The two examples of welfare analysis just described proceed within a partial equilibrium framework, treating each industry as negligible compared to the entire economy and regarding the marginal utility of money as approximately constant to each individual. Marshall's rather fragmentary remarks on optimal tax systems, income redistribution and the 'doctrine of maximum satisfaction' cannot be restricted in this way, and so raise serious unresolved analytical difficulties. On the other hand, his tax-subsidy argument was a valid counterexample to arguments that competition must lead to a social optimum, or that optimal indirect tax systems must involve uniform tax rates. It must also be borne in mind that utilitarian welfare economics was for Marshall only a first step towards a more evolutionary analysis of possible modes of improving the physical quality and the values and activities of mankind.

Bibliographic note: Marshall's treatment of welfare economics is to be found in (1920, pp. 18–19, 124–37, 462–76, 487–93). Ellis and Fellner (1943) is a good statement of the standard interpretation of the Marshall–Pigou tax-subsidy argument, emphasizing external effects. See also Bharadwaj (1972). On Marshall's treatment of compromise benefit see Whitaker (1986, pp. 186–8). Myint (1948) gives a useful general perspective on Marshall's welfare theory. Albon (1989) offers an intriguing insight into Marshall's attempt to apply welfare analysis to issues surrounding the British Post Office monopoly.

Interrelated Markets and Distribution Theory

Marshall was anxious to emphasize the interdependence of markets and introduced his treatment of joint and composite demand and supply largely for this purpose. A group of goods is jointly supplied if all are outputs of a single productive activity and jointly demanded if all are inputs. On the other hand, a particular good is compositely supplied or demanded if it is provided or acquired by several distinct productive activities. Marshall's formal treatment of joint demand and supply proceeded on the general assumption that the products involved were consumed or produced in fixed proportions, as did his related analysis of the 'derived demand' for any one of several jointly demanded inputs – 'derived' since the demand for such inputs is derived from

the demand for their joint product. The derived demand curve for a specific input can be constructed conceptually by supposing that its supply is perfectly elastic at an arbitrary price and that the markets for the output and all the inputs (including the specific input) adjust to equate quantity demanded to quantity supplied in each market. This gives a price quantity combination on the derived demand curve for the specific input. Other such combinations can be obtained by varying the arbitrarily chosen price and repeating the exercise, and so on. Marshall laid down four rules for inelasticity of derived demand. These were that the input should have no good substitutes, that the product it helps make should be inelastically demanded, that the input should account for only a small part of production costs, and that cooperating inputs should be inelastically supplied. Fixity of input ratios guaranteed the first condition, but the more general case was asserted rather than proven. The advantage of working with the derived demand curve for an input is that it permits a more transparent analysis of the effects of changes in the supply conditions of the singled-out input.

The prime example of joint demand is the demand for productive inputs, and Marshall's analysis of market interdependence was carried through more fully in this specific connection, the role of substitution among inputs receiving full acknowledgement. The principle of substitution ensured that input usage tended to be adjusted by firms so as to minimize the total production cost of any level of output. Thus, the value of the marginal product of an input (or the 'net product' as Marshall termed it) tended under competition to equal the unit price of the input. There has been some confusion about the relation between 'net product' and marginal product because the former allows usage of other inputs to adjust consequentially when the chosen input is increased while the latter does not. But, provided that the initial situation is cost-minimizing, the adjustment of other inputs (if small) has no effect on the change in output – an application of the envelope theorem. Marshall recognizes this explicitly (1920, p. 409n) and there is no good reason for refusing to classify him as a marginal productivity theorist.

Interdependence among input markets was further highlighted in the analysis of the competition of several industries for an input that is in temporarily or permanently fixed overall supply. A peculiarity of this last analysis was the insistence on excluding from the marginal cost of any industry the cost of bidding such fixed resources away from other uses. This is a perfectly legitimate application of the general envelope theorem: provided resource use is optimally adjusted, the marginal cost of increasing output will be the same whatever input or sub-group of inputs is increased. But Marshall's insistence on asymmetry where there is really symmetry can be accounted for only by his desire to legitimize, and extend to quasirent, the classical doctrine that rent is price determined rather than a pricedetermining element of cost.

Marshall's vision of market interdependence culminates in his treatment of income distribution, where he seeks to bring out the extents to which the interests of different factors of production are harmonious or conflicting. Distribution is determined by the interaction of the demands and supplies for the various inputs, the demands being essentially joint demands. Marginal productivity is a theory of input demand, not a complete theory of distribution, because the supplies of the various inputs cannot be viewed as fixed, at least in the long period. Indeed, in the long period the dominant influences on the prices of factors other than land are exerted by their supply conditions. The costs that then have to be met must ensure that various kinds of labour and capital continue to be replaced in their existing uses and quantities.

From an overall view 'The net aggregate of all the commodities produced is itself the true source from which flow the demand prices for all these commodities, and therefore for the agents of production used in making them' (1920, p. 536). This aggregate, 'the national dividend', is distributed among the factors of production. It is at once the

> aggregate net product of, and the sole source of payment for, all the agents of production within the country: it is divided up into earnings of labour; interest of capital; and lastly the producer's surplus, or rent, of land and of other differential advantages for production. It constitutes the whole of them, and

> the whole of it is distributed among them; and the larger it is, the larger, other things being equal, will be the share of each of them. (1920, p. 536)

The share going to any class of inputs will depend upon the need people have for its services: 'not the *total* need, but the *marginal* need' (p. 536, italics original). But a complicating influence for distribution theory, although one 'more full of hope for the future of the human race than any that is known to us' lies in the fact that 'highly paid labour is generally efficient and therefore not dear labour' (p. 510). Influenced by F.A. Walker, Marshall was a strong proponent of the 'economy of high wages' argument that high wages increase labour efficiency, not perhaps immediately, but cumulatively over time and perhaps over generations: effects that transcend simple theorizing in terms of static equilibrium.

All the different productive factors cooperating in production have a common interest in increasing the size of the pie to be shared, the national dividend or income, but each factor has a selfish interest in restrictive practices that increase its own share, even if they reduce the size of the pie slightly. A prime question of social policy for Marshall is how these divergent incentives can be reconciled: how combined action by various groups, such as unions, can be prevented from assuming forms that, while perhaps individually beneficial to any one group in isolation, are certainly mutually harmful if undertaken by all.

Marshall here enters into macroeconomic forms of argument, and it is indeed true that he did toy with the formal specification of macroeconomic models of growth and distribution (see Whitaker, 1975, vol. 2, pp. 305–16). But, with this exception, it should be emphasized that his treatment of market interdependence fell far short of a full theory of general equilibrium on Walrasian lines. Even when formalizing market interdependence in the mathematical appendix to *Principles* (1920, pp. 846–56), he simply treated the demand or supply of each commodity as a function of nothing but the price of the commodity itself. The links between the generation of income in factor markets and the expenditure of that income in product markets were left quite vague. Again, it must be recalled that the development of comprehensive fully articulated equilibrium theories was not Marshall's aim.

Bibliographic note: The key sections for Marshall's treatment of interrelated markets and distribution theory are (1920, pp. 381–54, 504–45, 660–67, 846–56). For general commentaries on Marshall's distribution theory see Stigler (1941), H.M. Robertson (1970), Whitaker (1974, 1988). On Marshall's treatment of labour supply see Walker (1974, 1975), Matthews (1990). On the economy of high wages see Petridis (1996).

Monopoly and Combination

Marshall's analysis of price and output determination by a profit-maximizing monopolist, and of the effects of taxing such a monopolist, followed the lead of Cournot. The concept of marginal revenue was implicit in the mathematical statement, but Marshall's chosen vehicle was geometrical. Curves of average revenue and cost, and of their difference, average net revenue, y, (all functions of the quantity sold, x) were superimposed on a grid of iso-profit hyperbolae of form $xy = constant$. Profit was maximized when the average net revenue curve touched the highest such iso-profit curve. Weighting consumer surplus into the maximand, as well as net revenue, gave rise to the welfare analysis of 'compromise benefit' already mentioned.

Monopoly analysis was applied to trades unions, with the use of the concept of the derived demand for an input. A union controlling a labour input for which derived demand is inelastic can certainly raise wages – not only the wage rate but the total wages received – although at the price of unemployment of some members. Whether such a monopolistic restriction can be sustained for long is more doubtful, as there will be pressures both to enter the union and to evade its grasp by the relocation or reorganization of production.

A more problematic question was whether 'labour's disadvantage in bargaining' meant that combined action by workers could raise wages, even without any restriction of labour supply. Marshall believed that it did, but emphasized that the result might be less capital accumulation

by non-workers, an outcome that could harm workers eventually.

The extremes of monopoly and competition were both covered by the theory of normal value, even though the competition might be more akin to later concepts of imperfect or monopolistic competition than to any ideal form of perfect competition. But 'normal action falls into the background, when Trusts are striving for the mastery of a large market' (1920, p. xiv). The incidents, tactics and alliances of oligopolistic conflict defied reduction to a simple general theory. They were to have been considered in the uncompleted second volume of *Principles* and were to some extent covered by *Industry and Trade.* The latter's treatment of entry-limiting behaviour by a 'conditional monopolist', who dominates the market but does not control entry, is of considerable interest in the light of much recent work on this class of problems.

Bibliographic note: Marshall's treatment of monopoly theory is to be found in (1879b, pp. 180–86; 1920, pp. 477–95, 856–8). For his views on trusts and conditional monopolies see (1890b; 1919, pp. 395–635, especially 395–422). For his views on trades unions see (1879b, pp. 187–213; 1892, pp. 362–402; 1920, pp. 689–722) and Petridis (1973). On 'labour's disadvantage in bargaining' see Hicks (1930). Liebhafsky (1955) summarizes the relevant arguments of *Industry and Trade.*

Monetary Theory

Marshall was in full command of previous British discussions of monetary issues, but not himself a major contributor to the development of monetary theory. His evidence before royal commissions in 1887 and 1899 showed an impressive mastery of monetary analysis, both domestic and international, and was minutely examined by successive generations of Cambridge students, serving for many years virtually as a textbook. But it was not until 1923, with the appearance of *Money Credit and Commerce*, that Marshall put forward his monetary views in a systematic way. By then these had not the novelty, nor he the vigour, to advance contemporary discussion.

Marshall's most important contribution to monetary theory was to place the overall demand for money in the context of individual choices as to the fraction of one's wealth to keep on hand as ready cash. This approach, set out clearly in a manuscript of the early 1870s (Whitaker 1975, Vol. I, pp. 164–77), was developed by Marshall's Cambridge successors, especially A.C. Pigou and F. Lavington, into what is termed the 'Cambridge k' approach. It laid the background for the treatment of the demand for money in J.M. Keynes (1936). On international monetary theory, Marshall espoused a form of purchasing power parity.

Marshall's name is particularly associated with his proposals for 'symmetalism', the use of a fixed-weight combination of gold and silver as the monetary base, and for indexed contracts based on a 'tabular standard of value', or price index, to be maintained by the government. The former was offered as an improvement on fixed-ratio bimetallism, of which he was never more than a lukewarm adherent.

Marshall had interesting, if fragmentary, insights into business fluctuations and general unemployment, which he viewed as temporary disequilibrium consequences of credit market dislocations. These spilled over into general coordination failures, with unemployment in one market spreading to others by reducing demand in cumulative fashion – the germ at least of the multiplier concept. On the other hand, Say's Law was maintained as an equilibrium truth of great importance. He saw the remedies for cyclical unemployment in the 'continuous adjustment of means to ends, in such a way that credit can be based on the solid foundation of fairly accurate forecasts', and in curbs on reckless inflations of credit that are 'the chief cause of all economic malaise' (1920, p. 710).

Bibliographic note: Marshall's monetary evidence is reproduced in J.M. Keynes (1926, pp. 3–195, 265–326). Other sources for his monetary views are Whitaker (1975, vol. 1, pp. 164–77), and Marshall (1887; 1923, pp. 12–97, 140–54, 225–33, 264–320). The standard treatment of Marshall's monetary views is Eshag (1963). For Marshall's views on business fluctuations see his

(1879b, pp. 150–57; 1885a; 1892, pp. 400–3; 1920, pp. 710–11; 1923, pp. 234–63). Also see Wolfe (1956), Laidler (1990).

International Trade

Marshall's major contribution to international trade theory was his well-known geometrical analysis of the equilibrium and stability of two-country trade by means of intersecting offer curves. Each country's offer curve indicated the number of 'bales' of home goods it was prepared to exchange for a specified number of bales of foreign goods, demand being elastic or inelastic as an increase in the latter caused the former to increase or decrease. Possibilities of multiple and unstable offer-curve intersections were noted. The offer curves themselves were taken as data, although complex readjustments of production and consumption underlay them. The need for a separate theory of international trade was justified, in classical vein, by the supposed international immobility of factors of production that remain mobile domestically.

The main purpose of this theoretical apparatus was to examine the effects of tariffs. A country might gain by selfishly exploiting its monopoly power through restricting trade, and would certainly gain if trading equilibrium occurred on an inelastic portion of the foreign offer curve. But Marshall came to doubt increasingly the transferability of this result to a multi-country case, although admitting that it might apply to an export tax on an exceptional commodity (like British steam coal) lacking close substitutes and incapable of being produced elsewhere.

A related attempt to construct a theoretical measure of the 'net benefit' a country gains from foreign trade, analogous to the measures of consumer and producer surplus, was not entirely satisfactory as the partial equilibrium context had clearly been transcended.

On matters of concrete trade policy for Britain, Marshall was a firm but cautious adherent of free trade, even unilateral free trade, but became increasingly concerned with the prospects for Britain's position in the world economy. The discussion in *Industry and Trade* of the links between foreign competition and domestic industrial organization and structure reflected this concern.

Bibliographic note: For Marshall's treatment of the theory of international trade by offer curves see Whitaker (1975, vol. 1, pp. 260–79; vol. 2, pp. 111–81), Marshall (1923, pp. 155–224, 330–60). For the net benefit measure see Whitaker (1975, vol. 1, pp. 379–81) and Marshall (1923, pp. 338–40). Commentaries on Marshall's theory are to be found in Viner (1937, pp. 527–92), Chipman (1965), Johnson and Bhagwati (1960) and Creedy (1990). For Marshall's views on trade policy and trends see Marshall (1919, pp. 1–177, 681–784; 1923, pp. 98–139, 201–24) and J.M. Keynes (1926, pp. 367–420).

A Brief Survey of Marshall's Writings with Suggestions for Further Reading

The first editions of Marshall's five books were (1879b), (1890a), (1892), (1919), (1923). *Economics of Industry* (1879b) had a new edition in 1881 and was reprinted with minor changes several times up to 1892. It is an important source for Marshall's views on distribution theory, trades unions and business fluctuations. His magnum opus, *Principles* (1890a), had new editions in 1891, 1895, 1898, 1907, 1910, 1916 and 1920. The title was changed to its final form (as in (1920)) in the fifth edition. *Principles* is the basic source for Marshall's views on the theories of value and distribution as well as his broader views on economics and social welfare. Since the rewritings between editions were substantial, the ninth variorum edition, edited by C.W. Guillebaud, Marshall's nephew (Guillebaud 1961), is essential for serious study. The first of its two volumes is a facsimile of the eighth edition of 1920. The second volume contains deleted passages from earlier editions, editorial notes, and various supporting documents. Users of the differently paginated Macmillan paperback edition of the eighth edition should note that all page references to the eighth edition given above must be located by using the table of correspondences appended to the paperback version.

Elements of the Economics of Industry (1892) had new editions in 1896 and 1899 and frequent reprintings. The last preface is dated 1907. It is essentially an abridgement of *Principles,* designed to replace *Economics of Industry,* a book that Marshall had come to despise, quite unjustifiably. *Elements* contains Marshall's fullest treatment of trades unions. *Industry and Trade* (1919) had new editions in 1919, 1920 and 1923 but only the first of these involved significant changes. Its three books deal with 'Some origins of present problems of industry and trade', 'Dominant tendencies of business organization', and 'Monopolistic tendencies: their relations to public well-being'. It adopts a largely historical and comparative approach and focuses on contemporary issues. Nevertheless, it contains many passages and insights of permanent interest and warrants closer attention by economists than it has received until recently. *Money Credit and Commerce* (1923) had only one edition and conveys Marshall's views on money, international trade and business fluctuations. Although blemished, it should not be dismissed.

An almost comprehensive annotated list of Marshall's occasional writings is found in Pigou (1925, pp. 500–8), which also reprints many of the texts of these writings. Guillebaud (1961) reproduces further occasional pieces. The 'Pure Theory' chapters (1879a), privately printed by Sidgwick, were first published in reprint form in 1930. A corrected and amplified version is included in Whitaker (1975). The two volumes of the latter also reproduce Marshall's unpublished early manuscripts, mainly from the 1870s, including several manuscript chapters from the abandoned volume on foreign trade. Marshall's important contributions to official enquiries are collected in J.M. Keynes (1926), a book that is supplemented by Groenewegen (1996).

The literature on Marshall's life and thought is too extensive to allow for more than a highlighting of some significant contributions. The splendid memorial essay by J.M. Keynes (1924) is not to be missed, although outdated on some points, nor is the charming memoir by Marshall's wife (Marshall 1944). Pigou (1925) includes fascinating vignettes by several of Marshall's colleagues and friends and Guillebaud (1971) gives a nephew's reminiscences. A major scholarly biography (Groenewegen 1995) covers Marshall's life and thought exhaustively, while a comprehensive three-volume edition of Marshall's correspondence (Whitaker 1996) provides much new information. Additional primary material on Marshall is provided in Raffaelli et al. (1995), where notes on Marshall's 1873 lectures to women students are reproduced, and Raffaelli (1994), where Marshall's essays on philosophical and psychological manuscripts from the late 1860s are reproduced and analyzed. See also Harrison (1963). Newspaper reports on public lectures Marshall gave during his years in Bristol are reproduced in Coase and Stigler (1969), Whitaker (1972) and Butler (1995).

Valuable overall assessments of Marshall are provided by Cannan (1924), Schumpeter (1941), Viner (1941), Shove (1942) and O'Brien (1981). Maloney (1985) studies Marshall's involvement in the professionalization of British economics. An extensive body of detailed analysis and criticism of Marshall's thought, mainly conducted in academic journals, continues to expand, with growing tributaries from Italy and Japan in particular. Wood (1982, 1996) assembles in eight volumes a somewhat miscellaneous collection of 239 pieces on Marshall, but standard bibliographic aids such as EconLit are recommended for a comprehensive search. The 1990 centenary of the publication of *Principles* produced two books of essays on Marshall (Whitaker 1990, McWilliams-Tullberg 1990) and several symposia on Marshall in economics journals. Samples of recent research can be found in Arena and Quéré (2003). On Marshall's social and behavioural views see Parsons (1931, 1932), Whitaker (1977) and Chasse (1984). For Marshall's views on socialism and trades unions see, respectively, McWilliams-Tullberg (1975) and Petridis (1973).

See Also

- ▶ Ceteris Paribus
- ▶ Consumer Surplus
- ▶ Demand Price
- ▶ External Economies
- ▶ Marshall, Mary Paley (1850–1944)

Selected Works

1872. Review of Jevons (1871). *Academy,* April. Reprinted in Pigou (1925).
1874. The future of the working classes. *The eagle.* Reprinted in Pigou (1925).
1876. On Mr. Mill's theory of value. *Fortnightly review,* April. Reprinted in Pigou (1925).
1879a. *The pure theory of foreign trade. The pure theory of domestic values.* Privately printed. Reprinted in 1930. London: London School of Economics, Scarce Works in Political Economy No. 1; and in amplified form in Whitaker (1975).
1879b. (With M.P. Marshall.) *The economics of industry,* 2nd edn. London: Macmillan, 1881.
1884. Where to house the London poor. *Contemporary review,* March. Reprinted in Pigou (1925).
1885a. How far do remediable causes influence prejudicially (a) the continuity of employment (b) the rate of wages? with four appendices. In *Report of proceedings and papers of the industrial remuneration conference*, ed. C. Dilke. London: Cassel. The important appendix on 'Theories and facts about wages' is also reproduced in Guillebaud (1961).
1885b. The present position of political economy: An inaugural lecture delivered at the Senate House Cambridge in February 1885. London: Macmillan. Reprinted in Pigou (1925).
1885c. On the graphic method of statistics. *Jubilee Volume,* a supplement to *Journal of the [London] Statistical Society.* Reprinted in Pigou (1925).
1887. Remedies for fluctuations of general prices. *Contemporary Review,* March. Reprinted in Pigou (1925).
1889. Cooperation. Presidential address to the 21st annual cooperative congress, Ipswich. Reprinted in Pigou (1925).
1890a. *Principles of economics, Volume One*. London: Macmillan.
1890b. Some aspects of competition. Presidential address to Section F of the British Association for the Advancement of Science. Reprinted in Pigou (1925).
1892. *Elements of economics of industry,* 3rd edn. London: Macmillan, 1899.
1983. On rent. *Economic Journal* 3, 74–90. Reprinted in Guillebaud (1961).
1897. The old generation of economists and the new. *Quarterly Journal of Economics* 11, 115–35. Reprinted in Pigou (1925).
1898. Distribution and exchange. *Economic Journal* 8, 37–59. Portions are reprinted in Pigou (1925) and Guillebaud (1961).
1902. A plea for the creation of a curriculum in economics and associated *branches of political science.* London: Macmillan. Reprinted in Guillebaud (1961).
1907. The social possibilities of economic chivalry. *Economic Journal* 17, 7–29. Reprinted in Pigou (1925).
1917. National taxation after the war. In *After-war problems,* ed. W. Dawson. London: George Allen and Unwin. Partly reproduced in Pigou (1925).
1919 *Industry and trade*, 4th edn. London: Macmillan, 1923.
1920 *Principles of economics: An introductory volume.* London: Macmillan. The eighth edition of Marshall (1890a).
1923. *Money, credit and commerce*. London: Macmillan.

Bibliography

Albon, R. 1989. Alfred Marshall and the consumers' loss from the British Post Office monopoly. *History of Political Economy* 21: 679–688.
Aldrich, J. 1996. The course of Marshall's theorizing about demand. *History of Political Economy* 28: 171–218.
American Economic Association. 1953. *Readings in price theory.* Homewood: Irwin.
Arena, R., and M. Quéré, eds. 2003. *The economics of Alfred Marshall: Revisiting Marshall's legacy.* Basingstoke/New York: Palgrave Macmillan.
Bharadwaj, K. 1972. Marshall on Pigou's wealth and welfare. *Economica* 39: 32–46.
Biswas, T. 1977. The Marshallian consumer. *Economica* 44: 47–56.
Butler, R.W. 1995. 'The economic condition of America': Marshall's missing speech at University College, Bristol. *History of Political Economy* 27: 405–416.
Cannan, E. 1924. Alfred Marshall, 1842–1924. *Economica* 4: 257–261.
Chasse, J. 1984. Marshall, the human agent and economic growth: Wants and activities revisited. *History of Political Economy* 16: 381–404.

Chipman, J. 1965. A survey of international trade: Part 2, the neoclassical theory. *Econometrica* 33: 685–760.
Chipman, J. 1990. Marshall's consumer's surplus in modern perspective. In Whitaker (1990).
Coase, R. 1975. Marshall on method. *Journal of Law and Economics* 18: 25–31.
Coase, R., and G. Stigler. 1969. Alfred Marshall's lectures on progress and poverty. *Journal of Law and Economics* 12: 181–226.
Cournot, A. 1838. *Mathematical principles of the theory of wealth*. Trans. N. Bacon. New York: Macmillan, 1897.
Creedy, J. 1990. Marshall and international trade. In Whitaker (1990).
Cunningham, W. 1892. The perversion of economic history. *Economic Journal* 2: 491–506.
Ellis, H., and W. Fellner. 1943. External economies and diseconomies. *American Economic Review* 33: 493–511.
Eshag, E. 1963. *From Marshall to Keynes: An essay on the monetary theory of the Cambridge school*. Oxford: Blackwell.
Friedman, M. 1949. The Marshallian demand curve. *Journal of Political Economy* 57: 463–495. Reprinted in Friedman, M. 1953. *Essays in positive economics.* Chicago: University of Chicago Press.
Frisch, R. 1950. Alfred Marshall's theory of value. *Quarterly Journal of Economics* 64: 495–524.
Gee, J. 1983. Marshall's views on 'short period' value formation. *History of Political Economy* 15: 181–205.
Groenewegen, P. 1995. *A soaring eagle: Alfred Marshall 1842–1924*. Aldershot/Brookfield: Edward Elgar.
Groenewegen, P., ed. 1996. *Official papers of Alfred Marshall: A supplement*. Cambridge: Cambridge University Press.
Guillebaud, C., ed. 1961. *Alfred Marshall: Principles of economics: Ninth (variorum) edition*, 2 vols. London: Macmillan.
Guillebaud, C. 1971. Some personal reminiscences of Alfred Marshall. *History of Political Economy* 3: 1–8.
Hague, D. 1958. Alfred Marshall and the competitive firm. *Economic Journal* 68: 673–690.
Harrison, R. 1963. Two early articles by Alfred Marshall. *Economic Journal* 73: 2–30.
Hicks, J. 1930. Edgeworth, Marshall and the indeterminateness of wages. *Economic Journal* 40: 215–231.
Jevons, W. 1871. *The theory of political economy.* London: Macmillan.
Johnson, H., and J. Bhagwati. 1960. Notes on some controversies in the theory of international trade. *Economic Journal* 70: 74–93.
Keynes, J.N. 1891. *The scope and method of political economy.* London: Macmillan.
Keynes, J. 1924. Alfred Marshall, 1842–1924. *Economic Journal* 34: 311–372. Reprinted with slight changes in Pigou (1925).
Keynes, J.M., ed. 1926. *Official papers of Alfred Marshall*. London: Macmillan.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.
Laidler, D. 1990. Alfred Marshall and the development of monetary economics. In Whitaker (1990).
Liebhafsky, H. 1955. A curious case of neglect: Marshall's industry and trade. *Canadian Journal of Economics* 21: 339–353.
McWilliams-Tullberg, R. 1975. Marshall's 'tendency to socialism'. *History of Political Economy* 7: 75–111.
McWilliams-Tullberg, R., ed. 1990. *Alfred Marshall in retrospect*. Aldershot/Brookfield: Edward Elgar.
Maloney, J. 1985. *Marshall, orthodoxy and the professionalisation of economics*. Cambridge: Cambridge University Press.
Marshall, M. 1944. *What I remember.* Cambridge: Cambridge University Press.
Matthews, R. 1990. Marshall and the labour market. In Whitaker (1990).
Mill, J.S. 1848. *Principles of political economy.* London: Parker. Many subsequent editions.
Myint, H. 1948. *Theories of welfare economics*. London: London School of Economics.
Newman, P. 1960. The erosion of Marshall's theory of value. *Quarterly Journal of Economics* 74: 587–600.
O'Brien, D.P. 1981. Alfred Marshall, 1842–1924. In *Pioneers of modern economics in Britain*, ed. D. O'Brien and J. Presley. London: Macmillan.
O'Brien, D. 1990. Marshall's work in relation to classical economics. In Whitaker (1990).
Opie, R. 1931. Marshall's time analysis. *Economic Journal* 41: 199–215.
Parsons, T. 1931. Wants and activities in Marshall. *Quarterly Journal of Economics* 46: 101–140.
Parsons, T. 1932. Economics and sociology: Marshall in relation to the thought of his time. *Quarterly Journal of Economics* 46: 316–347.
Petridis, A. 1973. Alfred Marshall's attitudes to and economic analysis of trade unions: A case of anomalies in a competitive system. *History of Political Economy* 5: 165–198.
Petridis, A. 1996. Brassey's law and the economy of high wages in nineteenth century economics. *History of Political Economy* 28: 583–606.
Pigou, A., ed. 1925. *Memorials of Alfred Marshall*. London: Macmillan.
Raffaelli, T. 1994. Alfred Marshall's early philosophical writings. *Research in the History of Economic Thought and Methodology: Archival Supplement* 4: 51–158.
Raffaelli, T. 2003. *Marshall's evolutionary economics*. London/New York: Routledge.
Raffaelli, T., E. Biagini, and R. McWilliams-Tullberg, eds. 1995. *Alfred Marshall's lectures to women: Some economic questions directly connected to the welfare of the laborer.* Aldershot/Brookfield: Edward Elgar.
Robbins, L. 1928. The representative firm. *Economic Journal* 38: 387–404.
Robertson, H. 1970. Alfred Marshall's aims and methods illustrated from his treatment of distribution. *History of Political Economy* 2: 1–65.

Robertson, D., P. Sraffa, and G. Shove. 1930. Increasing returns and the representative firm. *Economic Journal* 40: 76–116.
Schumpeter, J. 1941. Alfred Marshall's *Principles:* A semi-centennial appraisal. *American Economic Review* 31: 236–248.
Shove, G. 1942. The place of Marshall's *Principles* in the development of economic theory. *Economic Journal* 52: 294–329.
Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36: 535–550.
Stigler, G. 1941. *Production and distribution theories: The formative period*. New York: Macmillan.
Stigler, G. 1950. The development of utility theory. *Journal of Political Economy* 58 (307–27): 373–396.
Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3: 23–46. Reprinted in American Economic Association (1953).
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.
Viner, J. 1941. Marshall's economics in relation to the man and his times. *American Economic Review* 31: 223–235.
Walker, D. 1969. Marshall's theory of competitive exchange. *Canadian Journal of Economics* 2: 590–598.
Walker, D. 1974. Marshall on the long-run supply of labor. *Zeitschrift für die Gesamte Staatswissenschaft* 130: 691–705.
Walker, D. 1975. Marshall on the short-run supply of labor. *Southern Economic Journal* 41: 429–441.
Whitaker, J. 1972. Alfred Marshall: The years 1877 to 1885. *History of Political Economy* 4: 1–61.
Whitaker, J. 1974. The Marshallian system in 1881: Distribution and growth. *Economic Journal* 84: 1–17.
Whitaker, J., ed. 1975. *The early economic writings of Alfred Marshall, 1867–1890*, 2 vols. London: Macmillan.
Whitaker, J. 1977. Some neglected aspects of Alfred Marshall's economic and social thought. *History of Political Economy* 9: 161–197.
Whitaker, J. 1982. The emergence of Marshall's period analysis. *Eastern Economic Journal* 8: 15–29.
Whitaker, J. 1986. The continuing relevance of Alfred Marshall. In *Ideas in economics*, ed. R. Black. London: Macmillan.
Whitaker, J. 1988. The distribution theory of Marshall's principles. In *Theories of income distribution*, ed. A. Asimakopulos. Boston/Dordrecht/Lancaster: Kluwer Academic.
Whitaker, J., ed. 1990. *Centenary essays on Alfred Marshall*. Cambridge: Cambridge University Press.
Whitaker, J., ed. 1996. *The correspondence of Alfred Marshall, economist,* 3 vols. Cambridge: Cambridge University Press.
Wolfe, J. 1956. Marshall and the trade cycle. *Oxford Economic Papers* 8: 90–101.
Wood, J. 1982. *Alfred Marshall: Critical assessments,* 4 vols. London: Croom Helm.
Wood, J. 1996. *Alfred Marshall: Critical assessments, second series*, 4 vols. London: Routledge.

Marshall, Mary Paley (1850–1944)

G. Becattini

Keywords
Marshall, A.; Paley, M. P.

JEL Classifications
B31

British economist, born in Ufford (Nottinghamshire) on 24 October 1850; died in Cambridge 7 March 1944. Great-granddaughter of the great theologian William Paley, she was brought up in a strictly evangelical faith in Ufford, her father's vicarage. Thomas Paley, had taken a good degree in mathematics (33rd wrangler) in 1833 at Cambridge, and had been, for a period, a fellow of St John's College. Mary had one elder sister and two younger brothers.

In 1871, with a scholarship, she went up to Cambridge to complete her education with studies at university level. Under the whimsical chaperon Anne J. Clough (sister of the poet), and the teaching of a handful of young voluntary dons committed to the cause of higher education of women (among them Henry Sidgwick and Alfred Marshall), she took the Moral Sciences Tripos. She graduated, albeit informally, in 1874 (the first woman to achieve such a distinction in Cambridge) but the board of examiners (W.S. Jevons was among them) was so bitterly divided that in the certificate they recorded, very unusually, that she had received two votes for a first class and two for a second.

Shortly after her degree Mary Paley began to teach and to tutor female students in the newly opened Newnham Hall. In 1876, on request, she began to write a small economic textbook for Extension Lectures, that eventually became *The Economics of Industry* (1879). In the same year she became engaged to Alfred Marshall. They married in Ufford in July 1877. From that date onwards, till the death of Alfred Marshall in 1924, her life was essentially devoted, first in Bristol,

where they settled after marriage, then in Oxford (1883–4) and finally in Cambridge, to helping him in his scientific work and to saving him from all the normal nuisances of life.

For several decades Mary Paley Marshall taught and tutored female students of economics in Newnham college. A member of many associations (Charity Organization Society, Ethical Society, and so on) she participated in the founding group of the British Economic Association. After 1924 she became the first librarian of the newly founded Marshall Library, which she visited regularly until her 90th year. In 1928 the University of Bristol awarded her an honorary degree. She was a gifted amateur water colour painter and her posthumous Memoir, *What I Remember* (1947), shows glimpses of literary talent. Mary was not buried beside Alfred, but her ashes were scattered in the garden of her house.

Mary Paley Marshall's claims to be considered as an economist by herself are, strictly speaking, unassessable. Personally she signed only a few short notes in the early issues of the *Economic Journal,* which show a clear mind, a good style and a balanced judgement, but no more. Her only title to fame resides in the green-covered *Economics of Industry,* co-authored with Alfred Marshall. This small textbook, reprinted many times and translated into several foreign languages, was rated very highly by contemporaries. J.M. Keynes went so far as to say: 'It was, in fact, an extremely good book; nothing more serviceable for its purpose was produced for many years, if ever.' From the viewpoint of the development of economic analysis the book is relevant as a sort of half-way house between the *Principles* of J.S. Mill and the *Principles* of A. Marshall. Despite some hints to the contrary by J.M. Keynes, the respective positions (teacher and pupil) and ages (Alfred was older by eight years) suggest that Mary Paley's contribution was only secondary and subordinate.

Worthy of mention is the help she afforded Alfred Marshall in preparing and amending all his works. In a letter to John Neville Keynes, there is a hint of a substantial collaboration: 'My wife and I', writes Alfred Marshall, alluding to an article by J.L. Laughlin (*Quarterly Journal of Economics,* 1887), 'find it very hard to see Laughlin's points & perhaps we underrate the strength of his attack.'

Had it not been for the suffocating influence of Alfred, Mary Paley, with her clear mind, earnestness and strong will, would have become herself, we can confidently guess, an economist of repute and not, as is the case, a minor figure in the shadow of Alfred Marshall.

Selected Works

1879. (With Alfred Marshall.) *The economics of industry*, 2nd ed. London: Macmillan. 1881.
1896. Conference of women workers. *Economic Journal* 6(21), 107–109.
1947. *What I remember.* Cambridge: Cambridge University Press.

Bibliography

Clough, B.A., and J.P. Strachey. 1944. Mrs Alfred Marshall (Mary Paley) 1850–1944. *Cambridge Review* 65 (1597): 300.
Constable, W.G. 1960. *Art and economics in Cambridge*. The Eagle, April.
Keynes, J.M. 1944. Mary Paley Marshall 1850–1944. *Economic Journal* 54: 268–284. Reprinted in J.M. Keynes, *Essays in biography,* ed. G. Keynes. London: Rupert Hart-Davis. 2nd ed, 1951.
Keynes, J.N. (n.d.). Letter 1(58), kept in the Marshall Library, Cambridge.

Marshall–Lerner Condition

Murray C. Kemp

Keywords
Correspondence principle; International trade theory; Marshall–Lerner condition; Terms of trade; Transfer problem

JEL Classifications
F1

In the comparative-statical calculations of the simplest two-commodity barter theory of international trade, the outcomes invariably depend on the magnitude of the sum of the two import-demand elasticities, one relating to the country under study (the 'home' country) and the other to the rest of the world (collectively, the 'foreign' country); in particular, the response of a variable to a disturbance will be in one direction if the sum of elasticities is less than minus 1 and in the opposite direction if the sum is greater than minus 1. On the other hand, for some dynamic or 'disequilibrium' models of international trade it is a necessary and sufficient condition of local stability that the same sum of elasticities be less than minus 1. Let us define Δ as one plus the sum of the two elasticities of import demand. Then the so-called Marshall–Lerner condition requires that Δ be negative. Evidently the condition provides a link between the comparative-statics of international trade and some forms of trade dynamics. That such a link exists is, of course, the essence of Samuelson's correspondence principle.

Proceeding to a more detailed account of the Marshall–Lerner condition, let us suppose that the home country imports the first commodity, the foreign country the second; and let us denote by p the world price of the second commodity in terms of the first and by α and α^* parameters of the home and foreign economies, respectively. Then, in the absence of trade impediments and autonomous international transfers, we may write the general-equilibrium or *mutatis mutandis* home import-demand function as $\varphi(1/p, \alpha)$, the foreign import-demand function as $\varphi^*(p, \alpha^*)$ and the condition of world equilibrium as

$$\varphi(1/p, \alpha) - p\varphi^*(p, \alpha^*) = 0. \tag{1}$$

Suppose now that an initial equilibrium is disturbed by small changes in α and α^*. Differentiating (1) totally we obtain

$$\left(p^{-2}\varphi_{1/p} + \varphi^* + p\varphi_p^*\right) \mathrm{d}p = \varphi_\alpha \mathrm{d}\alpha - p\varphi_{\alpha*}^* \mathrm{d}\alpha^*$$

or, equivalently,

$$\Delta \mathrm{d}p \equiv (1 + \xi + \xi*)\mathrm{d}p = \left(\phi_\alpha \mathrm{d}\alpha - p\phi_\alpha^* \mathrm{d}\alpha^*\right)/\phi^* \tag{2}$$

where the subscripts indicate partial derivatives and where $\xi \equiv \varphi_{1/p}/(p\varphi)$ and $\xi^* \equiv p\varphi_p^*/\varphi^*$ are the price elasticities of home and foreign import demand, respectively.

On the other hand, we may consider the dynamic tâtonnement defined by the differential equation

$$\dot{p} \equiv \mathrm{d}p/\mathrm{d}t = f[\varphi^*(p, \alpha^*) - \varphi(1/p, \alpha)/p] \tag{3}$$

where f is a differentiable sign-preserving function of the world excess demand for the second commodity and t denotes time. Evidently (3) is a dynamic extension of (1). For the local stability of p at an equilibrium value it is necessary and sufficient that $\mathrm{d}f/\mathrm{d}p$ be negative in a sufficiently small neighbourhood of the equilibrium value. Now a little calculation shows that $\mathrm{d}f/\mathrm{d}p = f'\,\Delta\varphi^*/p$, where the prime indicates differentiation; moreover, since f is differentiable and sign-preserving, f' is necessarily positive sufficiently near the equilibrium value of p. For local stability, therefore, it is necessary and sufficient that Δ be negative.

By way of illustration we may consider the traditional 'transfer problem'. Identifying $\alpha = \alpha^*$ with the amount transferred from the foreign to the home country, in terms of the numeraire, Eq. (1) reduces to

$$\varphi(1/p, \alpha) - p\varphi^*(p, \alpha^*) - \alpha = 0.$$

and, if α is initially zero, Eq. (2) reduces to

$$\Delta \mathrm{d}p = \left(\varphi_\alpha - 1 + p\varphi_\alpha^*\right) \mathrm{d}\alpha/\varphi^*.$$

Evidently φ_α and $1 - p\varphi_\alpha^*$ are home and foreign marginal propensities to consume the first commodity. Thus, in stable systems, the terms of trade move in favour of the recipient of a small payment if and only if that country's marginal propensity to consume the imported commodity is less than the foreign country's marginal propensity to consume the same commodity.

The home and foreign economies have been specified in terms of the functions φ and φ^* only. Whether in any particular context the Marshall–Lerner condition is satisfied depends

on the structure imposed on φ and φ^* by the context. Evidently the appropriate structure will be quite different for economies with and without chronic unemployment and, more generally, for economies with and without internal market-clearing. It will also be quite different for rural and industrial economies, for growing and declining economies and for rich and poor economies.

While the inequality $\Delta < 0$ is now widely known as the Marshall–Lerner condition, the label is inappropriate in a context of dynamic analysis. For Alfred Marshall developed a quite different stability condition (see Marshall 1879, II; 1923, Appendix J; Samuelson 1947, pp. 266–7; Amano 1968; Kemp 1964, pp. 89–90); and Abba Lerner was not at all concerned with disequilibrium dynamics (see Lerner 1944).

Bibliography

Amano, A. 1968. Stability conditions in the pure theory of international trade: A rehabilitation of the Marshallian approach. *Quarterly Journal of Economics* 82 (2): 326–339.

Kemp, M.C. 1964. *The pure theory of international trade and investment*. Englewood Cliffs: Prentice-Hall.

Lerner, A.P. 1944. *The economics of control*. New York: Macmillan.

Marshall, A. 1879. *The pure theory of foreign trade*. Published privately. Reprinted with *The pure theory of domestic values*. London: London School of Economics and Political Science, 1949.

Marshall, A. 1923. *Money, credit and commerce*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Martin Stuart ('Marty') Feldstein (1939–)

Charles Yuji Horioka

Abstract

Martin Stuart ('Marty') Feldstein, currently George F. Baker Professor of Economics at Harvard University and President Emeritus of the National Bureau of Economic Research, Inc. (NBER), is an American economist who has made important contributions to public finance, macroeconomics, international economics, social insurance, health economics, the economics of national security, and many other fields of economics, trained a large number of prominent economists, served as President of the National Bureau of Economic Research for some 30 years, and served as President Ronald Reagan's chief economic advisor.

Keywords

Capital accumulation; Capital gains tax; Charitable giving; Council of Economic Advisors; Deadweight loss; Economics of national security; Euro; European Monetary Union; Feldstein; Feldstein–Horioka; Health economics; Health insurance; Home bias; Inflation; International capital flows; International capital mobility; Investment; National Bureau of Economic Research; Pensions; Public finance; Public pensions; Saving; Social insurance; Social security; Tax expenditures; Taxation; Unemployment compensation; Unemployment insurance

JEL Classifications

B31; D14; D22; E21; F21; F32; F33; F52; H20; H55; I13; J65

Introduction

Martin Stuart ('Marty') Feldstein was born in New York City on 25 November 1939. He graduated *summa cum laude* from Harvard College with a major in economics in 1961, and even though he was admitted by the Harvard Medical School, he opted to study at the University of Oxford on a Fulbright scholarship. He received a B. Litt. in 1963 and a Ph.D. in 1967, both in economics, from Oxford, but combined his interest in medicine and economics by writing a doctoral dissertation about how hospital costs could be reduced in a

government-run health system (the UK's National Health Service). He was a Fellow of Nuffield College, Oxford, from 1964 until 1967, before returning to the USA and to his alma mater. He became an Assistant Professor of Economics at Harvard University in 1967, and only 2 years later, at the age of 28, he was granted tenure and promoted to full professor, becoming one of the youngest full professors in Harvard's history.

Feldstein was an immensely popular teacher, teaching the introductory course in economics (which attracted as many as 1000 students) for 21 years, as well as graduate courses in macroeconomics, public finance, the economics of national security etc. He trained a huge cadre of students, such as Jeffrey D. Sachs and Lawrence H. Summers, who have gone on to pursue illustrious careers in academia and government.

Feldstein spent virtually his entire career at Harvard and is currently George F. Baker Professor of Economics there, but in addition to being an academic, Feldstein served as President of the National Bureau of Economic Research (NBER) for some 30 years from 1977 until 2008 (except in 1982–84), building it up to its current status as the premier US-based economic research organisation.

Feldstein has received honorary doctorates from several universities and is an Honorary Fellow of Nuffield College Oxford, a Corresponding Fellow of the British Academy, and a Fellow of the Econometric Society and the National Association for Business Economics.

Moreover, Feldstein has also made important contributions to economic policymaking as Chair of President Ronald Reagan's Council of Economic Advisors from 1982 until 1984, a member of President George W. Bush's Foreign Intelligence Advisory Board from 2007 until 2009, and a member of President Barack Obama's Economic Recovery Advisory Board from 2009 until 2011.

Feldstein currently serves on the Board of Directors of the Council on Foreign Relations, the Trilateral Commission, the Group of 30 (a Washington-based financial advisory body) and the National Committee on United States–China Relations. He also serves on the Council of Academic Advisors of the American Enterprise Institute and was formerly on the Boards of Directors of a number of major corporations including AIG, Eli Lilly and JP Morgan.

Well known for his ability to explain economic concepts clearly, Feldstein also frequently writes op-ed pieces for the *Wall Street Journal* and other newspapers and makes policy proposals about a variety of economic issues. In past years, he regularly co-authored such pieces with his economist wife Kathleen, who received a Ph.D. from the Massachusetts Institute of Technology (MIT).

With respect to Feldstein's research, he started out as a specialist in health economics, the field of his doctoral dissertation, but his interests gradually broadened to include social insurance more generally, public finance, macroeconomics, international economics, the European Monetary Union, the Chinese economy and the economics of national security, among others. Feldstein has written close to 400 scholarly articles on a broad range of topics, many of which show, via careful theoretical and empirical analysis, that government tax and transfer policies have important impacts on the economic behaviour of households and firms – in other words that households and firms respond to incentives and disincentives – and that the impact of tax and transfer policies cannot be gauged accurately unless one takes account of the behavioural responses that they induce.

In recognition of his impressive research achievements, in 1977 Feldstein was awarded the John Bates Clark Medal of the American Economic Association (AEA), which at the time was awarded every 2 years to the economist under the age of 40 who was judged to have made the greatest contribution to economic science, and he was elected President of the American Economic Association in 2004.

In 2007 Feldstein received the Bradley Prize from the Lynde and Harry Bradley Foundation; in 2011 he was included in the 50 Most Influential ranking of *Bloomberg Markets* magazine because, despite being a Republican himself, his staunch advocacy of scaling back tax expenditures (government spending through the tax code – i.e. tax exemptions, deductions or credits to select groups or specific activities) incurred the wrath of Republicans but garnered the widespread support

of Democrats [see Homan (2011) for more details]; and in 2012 he received the SIEPR Prize for Contributions to Economic Policy from the Stanford Institute for Economic Policy Research (SIEPR) of Stanford University.

Social Insurance

Feldstein has made significant contributions to a broad range of fields in economics, but one area in which he has truly revolutionised the thinking of economists and policymakers alike is in the area of social insurance [see Feldstein (2005b) for a useful overview]. Feldstein showed that, while existing social insurance programmes protect individuals from a variety of risks, they also distort individuals' behaviour in a variety of ways, thereby lowering saving, economic growth and welfare. For example, unemployment insurance protects individuals from the loss of income during unemployment spells, but at the same time causes individuals to search for new jobs for too long, induces them to save less and encourages them to take jobs that have a greater likelihood of seasonal and/or cyclical layoffs. Social security (public old-age pensions) protects individuals from poverty during old age and longevity risk, but at the same time induces individuals to reduce their saving and to retire too early. Health insurance protects individuals from the risk of being unable to afford needed medical care, but at the same time induces them to consume too much health care, to take inadequate precautions with their health and to save too little because they know that they will not have to bear the full cost of health care themselves.

Feldstein has done extensive research to show that social insurance programmes have indeed distorted individuals' behaviour in the expected ways. For example, Feldstein (1976b, 1978a) shows that unemployment insurance substantially increases the amount of temporary layoff unemployment, while Feldstein and Poterba (1984) show that the higher the replacement rate of unemployment benefits, the higher the reservation wage of unemployed workers, meaning that the longer will be the duration of unemployment.

Turning to Feldstein's work on the impact of social security [see Feldstein and Liebman (2002b) for a useful overview], Feldstein (1974), a very well-known article, analyses theoretically and empirically the impact of social security on household consumption and saving behaviour. As Feldstein notes, there are at least two offsetting ways in which a social security system can affect household saving. On the one hand, the introduction of a social security system will cause households to not feel as much need to save for retirement as they did before the introduction of such a system (the wealth replacement effect). On the other hand, the introduction of a social security system will induce households to retire earlier, lengthen their retirement span, and make it necessary for them to save more than before (the induced retirement effect). Thus, it is not clear *a priori* whether the introduction of a social security system will induce households to save more or less than before. However, Feldstein has shown, using a variety of data, that the wealth displacement effect dominates the induced retirement effect, as a result of which the net effect of the public pension system is to lower household saving and capital accumulation. For example, Feldstein (1974, 1982c, 1996b) shows, using time-series data for the USA, that social security has substantially increased personal consumption (and by implication has substantially reduced personal saving). [Leimar and Lesnoy (1982) pointed out that there was a programming error in the program used to calculate the social security wealth variable used in Feldstein (1974), but Feldstein (1982c) showed that correcting the error does not change the basic result, and moreover, Feldstein (1974) remains of seminal importance, even though the empirical results have been debated, because it focused attention on social security crowd-out issues.] Similarly, Feldstein (1977, 1980) estimates the magnitudes of both the wealth displacement effect and the induced retirement effect using cross-country data and shows that the former is larger in absolute

magnitude than the latter, as a result of which social security reduces private saving, on balance.

Turning to Feldstein's work on health insurance [see Feldstein (1995c) for a useful overview], Feldstein (1970) finds that the price of physicians' services rose much faster than the consumer price index after the introduction of Medicare and Medicaid in 1966, suggesting that these health insurance programmes increased the demand for physicians' services, thereby bidding up the price. Similarly, Feldstein (1973) estimates a demand function for health care using time series data on individual states of the USA and finds that health insurance has a significant impact on the demand for health care.

Moreover, after establishing that social insurance programmes have indeed distorted the behaviour of individuals as theory would predict, Feldstein has then gone on to make recommendations about how to reform these programmes so that they continue to protect individuals from the risks they are designed to protect them from, while at the same alleviating the distortions they cause, including the reductions in saving that they cause. For example, Feldstein proposes reforming the current unemployment insurance system by making benefits taxable (as they are now, due in large part to Feldstein's own efforts), reducing replacement rates and/or (better yet) introducing a more fundamental reform involving the introduction of unemployment insurance saving accounts earmarked to pay benefits when unemployment occurs [see, for example, Feldstein and Altman (2007)].

Similarly, Feldstein proposes a transition from the current pay-as-you-go social security system to a mixed system with a substantial investment-based (funded) component [see, for example, Feldstein (1995a, f, 1996a, 1997b, 1998, 2005c, 2009a), Feldstein and Samwick (1997, 1998a, b, 2002), Feldstein and Ranguelova (2001) and Feldstein and Liebman (2002a)]. Economists have traditionally felt that it is difficult, if not impossible, to make the transition from a pay-as-you-go social security system to a mixed system or a fully funded system because one generation will have to bear a so-called 'double burden', financing the social security benefits of its parents' generation as well as its own benefits, but Feldstein shows in the aforementioned papers that this is not an insurmountable problem.

Finally, Feldstein proposes a transition from the current pay-as-you-go comprehensive low-deductible low-coinsurance Medicare system (the US federal health care programme for those over the age of 65) to a health savings account (HSA) system that creates strong incentives to choose high deductibles and high coinsurance [see, for example, Feldstein (1971, 1999a), Feldstein and Gruber (1995) and Feldstein and Samwick (1997)]. Tax-advantaged health savings accounts have been available to taxpayers who have a high-deductible health plan (HDHP) since 2003, and these accounts could provide a model for Medicare reform.

A related paper (Feldstein 1995b) pertains to college scholarships, which are not a form of social insurance narrowly defined, but which have similar distortive effects on individual behaviour. Feldstein (1995b) shows that college scholarships are beneficial in that they enable the children of low-income families to attend college, but distortive in that expected parental contributions are higher and the amounts of financial assistance lower if parental assets are higher. This implies a hefty 'education tax rate' on saving and strongly discourages parents from saving.

Public Finance

Another major contribution Feldstein has made to economics is in the area of public finance. Feldstein has shown in a long series of papers that taxes distort the behaviour of households and firms, causing substantial deadweight losses, and that one needs to take full account of the behavioural response of households and firms to tax changes when estimating their impact. For example, Feldstein (1995a, d, 1999b) estimated the impact of the Tax Reform Act of 1986 using the compensated elasticity of taxable income with respect to changes in tax rates, which allows one

to take account of the fact that tax changes would affect not only labour supply, but also the forms of compensation, the investment of assets and the extent of spending on tax-deductible activities [see Feldstein and Feenberg (1996) for a similar analysis of the 1993 tax rate increases].

Feldstein has also argued that tax policies have had the effect of depressing saving and investment, thereby causing a substantial deadweight loss and lowering economic growth [see Feldstein (1995e) for a useful survey]. We have already discussed the saving-depressing effects of various social insurance programmes, but the tax system also has the effect of depressing saving. The combination of the corporate income tax and personal income taxes on interest, dividends and capital gains creates a substantial wedge between the pre-tax marginal product of capital and the net return received by individual savers (i.e. it raises the effective tax rate on capital income), which depresses saving, investment and economic growth.

Feldstein (1976a, 1978b, c, 1995e) points out that, when doing a welfare analysis, the deadweight loss arising from the current tax system must be compared to the deadweight loss of the tax system that replaces it, and Feldstein (1995e) shows that a shift from the current income tax to a consumption tax or a labour income tax will increase national saving and reduce the deadweight loss of the overall tax system.

Feldstein has also analysed the distortionary effects of specific taxes. For example, Feldstein and Yitzhaki (1978) show that capital gains taxes have discouraged the sale of equities and reduced tax revenue via the lock-in effect and that cutting the capital gains tax would *increase* tax revenues by encouraging investors to realise their capital gains.

As another example, Feldstein (1981) shows that the use of the 'historic cost' method of depreciation for tax purposes implies that higher inflation rates reduce the real value of future depreciation deductions and therefore raise the real net cost of investment, that the rise in the real net cost of investment can be substantial at recent inflation rates, and that allowing accelerated depreciation and valuing depreciation at replacement cost for tax purposes are alternate ways of alleviating the biases caused by the use of 'historical cost' depreciation.

As yet another example, Feldstein has written a series of papers on the impact of tax breaks for charitable contributions on the amount of charitable giving [see, for example, Feldstein and Clotfelter (1976)]. These papers find that the elasticity of charitable giving with respect to the price or net cost of giving is slightly greater than one, which implies that any reduction in price will increase the total contributions received by charitable organisations by more than the reduction in tax revenue.

Moreover, Feldstein has also shown that the distortions caused by the tax system are exacerbated by the interaction between an unindexed tax system and high inflation [see Feldstein (1983b) for a collection of a number of his papers on this topic]. For example, inflation increases households' nominal incomes and pushes them into higher and higher tax brackets, thereby raising their marginal tax rates (a phenomenon called 'bracket creep'). Similarly, Feldstein (1983b, 1995e) has shown that inflation distorts the measurement of capital income and raises the effective tax rate on the return to saving. For example, because nominal interest income and nominal capital gains (the nominal increase in asset values) are taxed at the personal level, interest income and capital gains are overstated and overtaxed when there is inflation. Furthermore, because depreciation is understated and inventory profits are overstated when there is inflation, corporate income is overstated and overtaxed (although the fact that nominal interest payments can be deducted from profits works in the opposite direction). For example, Feldstein and Summers (1978) show that, on balance, the high rates of inflation in the USA in the 1970s raised the effective tax rate on corporate income substantially, and Feldstein (1982a) shows that this, in turn, has substantially reduced business investment in the USA since the late 1960s. Moreover, Feldstein (1982b) shows that the interaction between the tax code and inflation has distorted not only the size of the capital stock, but also the allocation of capital between business capital and housing capital, leading to too little of the former and too much of the latter.

The Feldstein–Horioka Paradox or Puzzle

One of Feldstein's best known papers is Feldstein and Horioka (1980), his joint paper with Charles Yuji Horioka on international capital mobility [see also Feldstein (1983a) and Feldstein and Bacchetta (1991)]. In this paper, the authors regress the domestic investment rate on the domestic saving rate using cross-section data on OECD member countries for the 1960–74 period and find that the coefficient of the domestic saving rate (subsequently referred to as the 'saving retention coefficient') is significantly different from zero but not significantly different from one. If capital markets are fully integrated and capital flows freely across national borders in search of the highest return, this coefficient should be zero (at least in the case of small economies). The fact that it is not significantly different from one suggests (in the absence of legal and other restrictions on international capital flows) that 'home bias' is strong and that people strongly prefer to invest their saving in their home country. This result generated considerable interest and surprise, because many economists had assumed that capital had become perfectly mobile internationally, and it came to be known as the 'Feldstein–Horioka puzzle or paradox'. It spawned a voluminous literature trying to verify and/or explain the result [see Coakley et al. (1998) and Apergis and Tsoumas (2009) for surveys of this literature], was included in Obstfeld and Rogoff's (2001) list of the 'six major puzzles in international macroeconomics', is covered in many, if not most, textbooks in macroeconomics and international economics, and is one of the most frequently cited papers in international economics.

The findings of this paper are important for a number of reasons. First, they shed light on the true nature of the world capital market and on the extent to which capital markets are integrated. Second, they confirm that it is appropriate, at least as a first approximation, to study income distribution in general and tax incidence in particular using models that ignore international capital mobility. Third, they imply that the national return on domestic saving is approximately equal to the pre-tax domestic marginal product of capital, since such saving increases the domestic capital stock rather than either flowing abroad or replacing foreign investment at home.

The Economics of the European Monetary Union

A related topic on which Feldstein has also made an important contribution is the impact of the introduction of the European Monetary Union (EMU) and the euro. Feldstein did extensive research on this topic both before and after the introduction of the EMU and the euro, and has shown convincingly that the disadvantages of implementing a monetary union among a very heterogeneous group of countries have far outweighed the advantages thereof, and that the experiment was a dismal failure [see, for example, Feldstein (1997a, 2005a, 2012)].

The Economics of National Security

In recent years, Feldstein has become increasingly interested in the economics of national security, a topic of growing importance due to the rapid rise in terrorism and military conflicts. In addition to conducting research on various topics relating to national security, Feldstein has been advising graduate students on topics relating to national security; started a new graduate course on the Economics of National Security at Harvard shortly after the 11 September 2001 attacks; heads the Working Group on the Economics of National Security at the National Bureau of Economic Research; and founded the Economics of National Security Association (ENSA) in 2013.

To cite two examples of Feldstein's work in this area, Feldstein (2008) proposes ways of reforming the institutions for domestic counterterrorism (CT) in the USA and of strengthening cooperation among the CT activities of the USA and its allies, while Feldstein (2009b) draws lessons from the 1930s to make recommendations concerning economic and national security policies for the USA today, one of which is to increase the scale and funding of the Federal Bureau of Investigation

(FBI) and the Department of Homeland Security to prepare for the threats of domestic terrorism and cyber attacks that did not exist in the 1930s.

Economics Education

As an academic, Feldstein has made a substantial contribution not only to economics research but also to economics education. Perhaps most importantly, he has trained numerous prominent economists who went on to occupy the top echelons of academia and government, among them Alan J. Auerbach (Robert D. Burch Professor of Economics and Law; Director of the Burch Center for Tax Policy and Public Finance; formerly Chair of the Economics Department, University of California, Berkeley; and formerly Deputy Chief of Staff of the U.S. Joint Committee on Taxation), Raj Chetty (professor at Harvard University; a 2012 MacArthur Fellow; and recipient of the 2013 John Bates Clark Medal), David T. Ellwood (professor and Dean of the John F. Kennedy School of Government at Harvard University and formerly Assistant Secretary for Planning and Evaluation at the U.S. Department of Health and Human Services), Douglas W. Elmendorf (Director of the Congressional Budget Office; formerly Deputy Assistant Secretary for Economic Policy at the US Department of the Treasury; and formerly Senior Fellow at the Brookings Institution), Charles Yuji Horioka (Research Professor at the Asian Growth Research Institute in Kitakyushu City, Japan; previously Vea Family Professor of Technology and Evolutionary Economics Centennial at the University of the Philippines, Diliman; and recipient of the 2001 Japanese Economic Association-Nakahara Prize), R. Glenn Hubbard (professor and Dean of the Business School at Columbia University; formerly Deputy Assistant Secretary for Tax Analysis at the US Department of the Treasury; and formerly Chairperson of the President's Council of Economic Advisors), Steven N. Kaplan (Neubauer Family Distinguished Service Professor of Entrepreneurship and Finance at the University of Chicago and one of the top twelve Business School professors in the country according to *Business Week*), Jeffrey B. Liebman (Malcolm Wiener Professor of Public Policy, John F. Kennedy School of Government, Harvard University, and formerly Acting Deputy Director for Policy in the US Office of Management and Budget), Lawrence B. Lindsey (Chief Executive Officer of the Lindsey Group; formerly Director of the National Economic Council; and formerly member of the Board of Governors of the Federal Reserve System), Jose Piñera (Distinguished Senior Fellow at the Cato Institute; formerly Chile's Secretary of Labor and Social Security and Secretary of Mining; and architect of Chile's private pension system), James M. Poterba (professor and formerly Head of the Economics Department at the Massachusetts Institute of Technology; Feldstein's successor as President of the National Bureau of Economic Research; and formerly a member of the President's Advisory Panel on Federal Tax Reform), Harvey S. Rosen (professor and formerly Chair of the Economics Department at Princeton University; formerly Deputy Assistant Secretary for Tax Analysis at the US Department of the Treasury; and formerly member and Chairperson of the President's Council of Economic Advisors), Jeffrey D. Sachs (Professor and Director of the Earth Institute at Columbia University; formerly Professor at Harvard University; advisor to many Latin American and post-Communist economies; and Special Adviser to two successive United Nations Secretaries-General on the Millennium Development Goals), Andrew A. Samwick (Sandra L. and Arthur L. Irving ′72a, P′10 Professor of Economics; Director of the Nelson A. Rockefeller Center for Public Policy and the Social Sciences, Dartmouth College; and formerly Chief Economist of the President's Council of Economic Advisors), Joel Slemrod (Professor, Economics Department Chair and Director of the Office of Tax Policy Research at the University of Michigan, Ann Arbor; and a member of the Congressional Budget Office Panel of Economic Advisers), Lawrence H. Summers (Professor and formerly President of Harvard University; recipient of the 1993 John Bates Clark Medal; and formerly Secretary of the Treasury, Chief Economist of the World Bank and Chairperson of the President's

National Economic Council); and this is a very incomplete list.

Furthermore, Feldstein taught the introductory course in economics at Harvard for 21 years (from 1984 until 2005), and this course was the most popular course on campus when he was teaching it, with more than 1000 students enrolled in it in some years.

The National Bureau of Economic Research

Feldstein contributed significantly to the development of the field of economics by serving as the President of the National Bureau of Economic Research, Inc. (NBER), the US's leading non-profit, non-partisan economic research organisation, for some 30 years. He became President in 1977 at the tender age of 37, and served in this capacity until he retired in 2008 (except from 1982 to 1984, when he was Chairperson of the Council of Economic Advisors under President Reagan). The NBER expanded and prospered enormously during his many years at the helm and now counts more than 1300 leading scholars from throughout North America as Research Associates. Moreover, Feldstein organised the NBER into Research Programs (of which there are now 20) and Working Groups (of which there are now 15), created the NBER Working Paper series (dubbed 'yellow jackets' because of their distinctive yellow covers), inaugurated the NBER Summer Institute (a three-week gathering of applied economists that now draws 2400 participants a year), moved the headquarters to a modern office building near the Harvard campus in the direction of MIT, and fostered ties with the Centre for Economic Policy Research (its European counterpart), the Chinese Academy of Sciences etc. [see Warsh (2008) for more details]. One indication of the NBER's prominence is that 24 Nobel Prize winners in Economics and 13 past Chairs of the President's Council of Economic Advisers have been researchers at the NBER.

Feldstein is still active at the NBER as President Emeritus and Director of the Working Group on the Economics of National Security. Moreover, to honour Feldstein's outstanding tenure as President, the NBER inaugurated the Martin Feldstein Lecture, an annual lecture by an outstanding economist, in 2009.

Economic Policymaking

In addition, Feldstein has been deeply involved in economic policymaking for much of his career. In 1982–84, he served for 2 years as Chairperson of the Council of Economic Advisors under President Reagan and spoke out strongly about the dangers of ballooning federal government deficits, even though this led to conflicts with other members of the Reagan Administration and his advice was not always heeded. Feldstein has continued speaking out about the dangers of ballooning government deficits, and in recent years his staunch advocacy of scaling back tax expenditures as a way of trimming federal government deficits has incurred the wrath of Republicans, who are against any increases in tax revenues, but has garnered the widespread support of Democrats.

With respect to other policies upon which Feldstein has had an influence, he was instrumental in cutting the capital gains tax [see Loungani (2004)] and making unemployment compensation benefits taxable, and has long been an avid advocate of social security reform, was instrumental in implementing a tax reform that made social security benefits taxable in the case of high-income households in 1983, and was one of the main driving forces behind former President George W. Bush's initiative of partial privatisation of the social security system in 2005. President Bush's initiative ultimately failed, but much earlier (in 1980), one of Feldstein's disciples, Jose Piñera, successfully privatised Chile's pension system and converted it into a fully funded system based on personal retirement accounts as Secretary of Labor and Social Security.

Thus Feldstein has made remarkable contributions not only to the development of the field of economics, but also to the conduct of economic policy as a researcher, research administrator, policy advisor, educator and journalist.

See Also

- Budget Deficits
- Charitable Giving
- Euro
- Excess Burden of Taxation
- Generational Accounting
- Greek Crisis in Perspective: Origins, Effects and Ways-Out
- Health Econometrics
- Implicit Contracts
- International Real Business Cycles
- Inventory Investment
- Investment (Neoclassical)
- Labour Supply
- Laffer Curve
- Macroeconomic Effects of International Trade
- National Bureau of Economic Research
- Neutrality of Money
- Public Finance
- Ricardian Equivalence Theorem
- Social Insurance and Public Policy
- Social Security in the United States
- Tax Incidence
- Unemployment Insurance

Selected Works

Feldstein, M.S. 1970. The rising price of physician's services. *Review of Economics and Statistics* 52(2): 121–133.

Feldstein, M.S. 1971. A new approach to national health insurance. *The Public Interest* 23: 93–105.

Feldstein, M.S. 1973. The welfare loss of excess health insurance. *Journal of Political Economy* 81(2), part I: 251–280.

Feldstein, M.S. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82(5): 905–926.

Feldstein, M.S. 1976a. On the theory of tax reform. *Journal of Public Economics* 6(1–2): 77–104.

Feldstein, M.S. 1976b. Temporary layoffs in the theory of unemployment. *Journal of Political Economy* 84(5): 937–957.

Feldstein, M.S. 1977. Social security and private savings: International evidence in an extended life cycle model. In *The economics of public services* (International Economic Association Conference Volume), ed. M. Feldstein and R. Inman, 174–206. London: Macmillan.

Feldstein, M.S. 1978a. The effect of unemployment insurance on temporary layoff unemployment. *American Economic Review* 68(5): 834–846.

Feldstein, M.S. 1978b. The rate of return, taxation and personal savings. *Economic Journal* 88(351): 482–487.

Feldstein, M.S. 1978c. The welfare cost of capital income taxation. *Journal of Political Economy* 86(2): S29–S51.

Feldstein, M.S. 1980. International differences in social security and saving. *Journal of Public Economics* 14(2): 225–244.

Feldstein, M.S. 1981. Adjusting depreciation in an inflationary economy: Indexing versus acceleration. *National Tax Journal* 34(1): 29–43.

Feldstein, M.S. 1982a. Inflation, tax rules and investment: Some econometric evidence. *Econometrica* 50(4): 825–862.

Feldstein, M. 1982b. Inflation, tax rules and the accumulation of residential and nonresidential capital. *Scandinavian Journal of Economics* 84(2): 293–311.

Feldstein, M.S. 1982c. Social security and private saving: Reply. *Journal of Political Economy*, 90(3): 630–642.

Feldstein, M.S. 1983a. Domestic saving and international capital movements in the long run and the short run. *European Economic Review* 21(1–2): 129–151.

Feldstein, M.S. 1983b. *Inflation, tax rules, and capital formation* (National Bureau of Economic Research Books). Chicago: University of Chicago Press.

Feldstein, M.S. 1995a. Behavioral responses to tax rates: Evidence from the Tax Reform Act of 1986. *American Economic Review* 85(2): 170–174.

Feldstein, M.S. 1995b. College scholarship rules and private saving. *American Economic Review* 85(3): 552–566.

Feldstein, M.S. 1995c. The economics of health and health care: What have we learned? What have I learned? *American Economic Review* 85(2): 28–31.

Feldstein, M.S. 1995d. The effect of marginal tax rates on taxable income: A panel study of the 1986 Tax Reform Act. *Journal of Political Economy* 103(3): 551–572.

Feldstein, M.S. 1995e. Fiscal policies, capital formation, and capitalism. *European Economic Review* 39(3–4): 399–420.

Feldstein, M.S. 1995f. Would privatizing social security raise economic welfare? *NBER Working Paper No. 5281*, National Bureau of Economic Research.

Feldstein, M.S. 1996a. The missing piece in policy analysis: Social security reform. *American Economic Review* 86(2): 1–14.

Feldstein, M.S. 1996b. Social security and saving: New time series evidence. *National Tax Journal* 49(2): 151–164.

Feldstein, M.S. 1997a. The political economy of the European Economic and Monetary Union: Political sources of an economic liability. *Journal of Economic Perspectives* 11(4): 23–42.

Feldstein, M.S. 1997b. Public policy and financial markets: Privatizing social security. *Journal of Finance* 52(3): 1182–1184.

Feldstein, M.S. 1998. Transition to a fully funded pension system: Five economic issues. In *Redesigning social security,* ed. H. Siebert, 299–315. Tubingen: Mohr (Siebeck).

Feldstein, M.S. 1999a. Prefunding medicare. *American Economic Review* 89(2): 222–227.

Feldstein, M.S. 1999b. Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics* 81(4): 674–680.

Feldstein, M.S. 2005a. The euro and the stability pact. *Journal of Policy Modeling* 27(4): 421–426.

Feldstein, M.S. 2005b. Rethinking social insurance. *American Economic Review* 95(1): 1–24.

Feldstein, M.S. 2005c. Structural reform of social security. *Journal of Economic Perspectives* 19(2): 33–55.

Feldstein, M.S. 2008. Designing institutions to deal with terrorism in the United States. *American Economic Review* 98(2): 122–126.

Feldstein, M.S. 2009a. Reducing the risk of investment-based social security reform. In *Social security policy in a changing environment,* ed. J.R. Brown, J.B. Liebman, and D.A. Wise, 201–218. Chicago: University of Chicago Press.

Feldstein, M.S. 2009b. Economic conditions and U.S. national security in the 1930s and today. *NBER Working Paper No. 15290*, National Bureau of Economic Research.

Feldstein, M.S. 2012. The failure of the Euro: The little currency that couldn't. *Foreign Affairs* 91(1 (January–February).

Feldstein, M.S. and D. Altman 2007. Unemployment insurance savings accounts. In *Tax policy and the economy*, ed. J.M. Poterba, vol. 21, 35–64. Cambridge, MA: MIT Press.

Feldstein, M.S., and P. Bacchetta. 1991. National saving and international investment. In *National saving and economic performance,* ed. B.-D. Bernheim and J.B. Shoven, 201–226. Chicago: University of Chicago Press.

Feldstein, M.S., and C. Clotfelter. 1976. Tax incentives and charitable contributions in the United States: A microeconometric analysis. *Journal of Public Economics* 5(1–2): 1–26.

Feldstein, M.S., and D. Feenberg. 1996. The effect of increased tax rates on taxable income and economic efficiency: A preliminary analysis of the 1993 tax rate increases. In *Tax policy and the economy,* ed. J.M. Poterba, vol. 10, 89–118. Cambridge, MA: MIT Press.

Feldstein, M.S. and J. Gruber. 1995. A major risk approach to health insurance reform. In *Tax policy and the economy*, ed. J.M. Poterba, vol. 9, 103–130. Cambridge, MA: MIT Press.

Feldstein, M.S. and C. Horioka. 1980. Domestic saving and international capital flows. *Economic Journal* 90(358): 314–329.

Feldstein, M.S., and J.B. Liebman. 2002a. The distributional effects of an investment-based social security system. In *The distributional aspects of social security and social security reform,* ed. M.-S. Feldstein and J.B. Liebman, 263–322. Chicago: University of Chicago Press.

Feldstein, M.S., and J.B. Liebman. 2002b. Social security. In *Handbook of public economics,*

ed. A.J. Auerbach, and M.S. Feldstein, 2245–2324. Amsterdam: Elsevier Science, North-Holland.
Feldstein, M.S., and J. Poterba. 1984. Unemployment insurance and reservation wages. *Journal of Public Economics* 23(1–2): 141–167.
Feldstein, M.S., and E. Ranguelova. 2001. Individual risk in an investment-based social security system. *American Economic Review* 91(4): 1116–1125.
Feldstein, M.S., and A. Samwick. 1997. The economics of prefunding social security and Medicare benefits. In *NBER macroeconomic annual 1997,* ed. B.S. Bernanke and J.J. Rotemberg, 115–148. Cambridge, MA: MIT Press.
Feldstein, M.S., and A. Samwick. 1998a. Potential effects of two percent personal retirement accounts. *Tax Notes* 79(5): 615–620.
Feldstein, M.S., and A. Samwick. 1998b. The transition path in privatizing social security. In *Privatizing social security,* ed. M.S. Feldstein, 215–260. Chicago: University of Chicago Press.
Feldstein, M.S., and A. Samwick. 2002. Potential paths of social security reform. In *Tax policy and the economy 2001*, ed. J.M. Poterba, vol. 16, 181–224. Cambridge, MA: MIT Press.
Feldstein, M.S., and L. Summers. 1978. Inflation, tax rules, and the long term-interest rate. *Brookings Papers on Economic Activity* 9(1): 61–110.
Feldstein, M.S., and S. Yitzhaki. 1978. The effects of the capital gains tax on the selling and switching of common stock. *Journal of Public Economics* 6(1): 17–36.

Acknowledgments I am grateful to James Poterba and an anonymous referee for their detailed and helpful comments.

Bibliography

Apergis, N., and C. Tsoumas. 2009. A survey of the Feldstein–Horioka puzzle: What has been done and where we stand. *Research in Economics* 63(2): 64–76.
Coakley, J., F. Kulasi, and R. Smith. 1998. The Feldstein–Horioka puzzle and capital mobility: A review. *International Journal of Finance and Economics* 3: 169–188.
Homan, T.R. 2011. Feldstein on taxes sways Democrats more than fellow Republicans. Bloomberg Markets Magazine, September 7.
Leimar, D.R., and S.D. Lesnoy. 1982. Social security and private saving: New time-series evidence. *Journal of Political Economy* 90(3): 606–629.
Loungani, P. 2004. (People in economics) Getting there first: An economist's lifelong study of the effects of taxes and social insurance. *Finance and Development* (March): 4–7.
Obstfeld, M., and K.S. Rogoff. 2001. The six major puzzles in international macroeconomics: Is there a common cause? In *NBER macroeconomics annual 2000*, vol. 15, 339–412. Cambridge, MA: MIT Press.
Warsh, D. 2008. *Economic principals: He changed economics*. Blog posted on 27 July 2008. http://www.economicprincipals.com/issues/2008.07.27/328.html. Accessed 18 Mar 2015.

Martin, Henry (Died 1721)

Douglas Vickers

Henry Martin (Martyn), eldest son of Edward Martyn of Upham, Wiltshire, owes his place in the history of economics to his participation in the debates surrounding the Treaty of Commerce between England and France proposed at Utrecht in 1713. He did not pursue the legal career for which he had been educated, but he accepted, as a reward for his contributions to the trade debates, an appointment as Inspector-General of Exports and Imports. He died at Blackheath in March 1721.

Martin was the principal author (together with Joshua Gee, among many others) of the periodical *The British Merchant, or Commerce Preserved* which appeared twice a week between August 1713 and July 1714. It strenuously argued the case for continued protectionism and the support of the British manufacturing trades against the free-trade and reciprocal tariff clauses of the proposed Treaty, and it was undoubtedly due to Martin's writing that the Treaty was ultimately rejected.

Martin was also influential in the publication of the *Spectator*, and appears to have written issue No. 180 and to have contributed heavily to, or written, Nos. 200 and 232. These contributions

present the popular argument that the 'riches [of the state] must increase or decrease in proportion to the number and riches of the people'. Increased population would lead to increased consumption expenditures and would increase land values and thereby taxation revenue. Arguing that 'the wages of labourers make the greatest part of the price of everything', it was claimed that lower wages would reduce manufacturing prices and 'increase foreign markets'. Employment would be provided for a larger number of workers, particularly if, in the manner of William Petty's example of watch making, output volumes were increased by a division of labour and specialization of production (*Spectator* No. 232). For these reasons it was argued that the poor should be employed rather than made the object of indiscriminate charity.

See Also

► Mercantilism

Bibliography

Bond, D.F., ed. 1965. *The Spectator. 5 vols*. Oxford: Clarendon Press.

King, C. 1721. *The British merchant, or Commerce preserved*. 3 Vols. London. Reprinted, New York: Kelley, 1968.

Martineau, Harriet (1802–1876)

Robert W. Dimand and Evelyn L. Forget

Keywords

Classical political economy; Laissez-faire; Marcet, J.; Martineau, H.; Property rights; Slavery

JEL Classifications

B31

Harriet Martineau, the best-selling popularizer of classical political economy, was born in Norwich, England, on 12 June 1802, the sixth of eight children of Thomas Martineau, a Unitarian textile manufacturer, and Elizabeth Rankin Martineau. She was educated at home, except that from 1813 to 1815 she studied French, Latin, and English composition at a school run by the Reverend Isaac Perry. Her early writings were religious, beginning with an article on 'Female Writers on Practical Divinity' for the *Monthly Repository*. After her father's death in 1826, she became engaged, but her fiancé died before they could be married. She remained single for the rest of her life. Investment losses in 1829 forced her to support herself by writing: William Johnson Fox's *Monthly Review* hired her as a book reviewer for 15 pounds a year, and when the Central Unitarian Association offered prizes for essays to convert Catholics, Jews, and Muslims, Martineau won all three prizes, for 15 guineas each. These prizes enabled her to visit her brother in Dublin in 1831. While there, she planned the series *Illustrations of Political Economy*, stories that would expound (especially to the working classes) the principles of classical political economy, to which she had been introduced by Jane Marcet's *Conversations on Political Economy* (1816) and James Mill's *Elements of Political Economy* (1821). The first of the 34 tales of political economy, of the Poor Laws, and of taxation was published in February 1832 by Charles Fox (brother of William Fox), and distributed by the Society for the Diffusion of Useful Knowledge. By 10 February, the first printing of 1,500 copies had all been sold, and a second printing of 5,000 copies ordered.

The success of *Illustrations of Political Economy* (1832–4, 2004a) made Martineau a celebrity: Henry Brougham, the Lord Chancellor, provided her with private papers on the impending reform of the Poor Law, while Robert Owen attempted, without success, to convert her to socialism. Although poor health interrupted her work several times, Harriet Martineau refused offers of government pensions from Lord Grey in 1835, Lord Melbourne in 1841, and W.E. Gladstone in 1873, lest her independence be compromised. Instead, her friends raised funds for an annuity

for her in 1843, and, when her health permitted she worked with great diligence, writing more than 1,600 articles for the *Daily News* from 1852 to 1866. Although her income was modest, she was a philanthropist, founding a building society among other charitable projects.

Harriet Martineau disclaimed originality for *Illustrations of Political Economy*, insisting that its purpose was didactic, making established principles better known. Her presentation of the principles was original, and her work stood out for its recognition of women as rational economic agents (1985). In her *Illustrations*, Martineau upheld laissez-faire and property rights. In later life, Martineau endorsed married women's property rights, condemned slavery as an illegitimate form of property, gave qualified support to workers' cooperatives, and accepted the need for state intervention in certain very limited circumstances. Her lectures at working men's institutions stressed education rather than state intervention as the remedy for most social ills. Her subsequent writings on America and slavery (1837, 2002) demonstrated that she was an adept economic analyst, not just a popularizer. For example, Martineau recognized the limited demand for the services of prostitutes in the South as evidence of the sexual exploitation of slaves, and identified the inability of slaveholders to make credible long-term commitments to their slaves as a source of inefficiency in slave agriculture (Levy 2003). Martineau's *Society in America* (1837) emphasized the incompatibility of slavery and of the legal, political, and economic position of American women (lacking votes and, if married, property rights) with America's founding rhetoric of liberty. Her study of the United States was accompanied by her *How to Observe Morals and Manners* (1838), a methodological manual on comparative sociology and ethnography. In 1853, Martineau published a translation and abridgement of the pioneer sociologist Auguste Comte's *Philosophie positive*, an adaptation that pleased Comte so much that he had it translated back into French.

In 1855, anticipating death from heart disease, Martineau wrote her autobiography for posthumous publication (1877), but she lived until 27 June 1876.

See Also

- ▶ British Classical Economics
- ▶ Marcet, Jane Haldimand (1769–1858)
- ▶ Slavery

Selected Works

1832–4. *Illustrations of political economy, taxation, poor law and paupers*, 13 vols. London: C. Fox; repr. with new introduction by C. Franklin Bristol: Thoemmes, 2001.

1837. *Society in America*. London: Saunders and Otley. Abridged edn, ed. S.M. Lipset, New Brunswick: Transaction, 1981.

1838. *How to observe morals and manners*. London: C. Knight. Repr. 1989, ed. M.R. Hill. New Brunswick: Transaction.

1877. *Harriet Martineau's autobiography*, 2 vols. London: Smith, Elder; repr. London: Virago, 1983.

1985. *Harriet Martineau on women*, ed. G.G. Yates. New Brunswick: Rutgers University Press.

2002. *Writings on slavery and the American civil war by Harriet Martineau*, ed. D.A. Logan. DeKalb: Northern Illinois University Press.

2004a. *Illustrations of political economy: Selected tales*, ed. D.A. Logan. Peterborough: Broadview Press.

2004b. *Harriet Martineau's writings on the British Empire*, 5 vols, ed. D.A. Logan. London: Pickering & Chatto.

2005. *Harriet Martineau: Writings on British history and military reform*, 6 vols, ed. D.A. Logan. London: Pickering & Chatto.

2006. *The collected letters of Harriet Martineau*, 5 vols, ed. D.A. Logan. London: Pickering & Chatto.

Bibliography

Henderson, W. 1992. Harriet Martineau or when political economy was popular. *History of Education* 21: 383–403.

Levy, D. 2003. Taking Harriet Martineau's economics seriously. In *The status of women in classical economic thought*, ed. R. Dimand and C. Nyland. Cheltenham: Edward Elgar.

Logan, D.A. 2002. *The hour and the woman: Harriet Martineau's somewhat remarkable' life*. DeKalb: Northern Illinois University Press.
Marcet, J. 1816. *Conversations on political economy*. London: Longman. 3rd ed., 1818.
Mill, J. 1821. *The elements of political economy*. London: Baldwin, Cradock & Joy.
Orazem, C. 1999. *Political economy and fiction in the early writings of Harriet Martineau*. Frankfurt: Peter Lang.
Webb, R.K. 1960. *Harriet Martineau, a radical victorian*. London: Columbia University Press.

Martingales

Alan F. Karr

JEL Classifications
C0

A martingale is a mathematical model of a fair game, or of some other process that is incrementally random noise. The term, which also denotes part of a horse's harness or a ship's rigging, refers in addition to a gambling system in which every losing bet is doubled; it was introduced into probability theory by J.L. Doob. Among stochastic processes, martingales have particular constancy properties with respect to conditioning. The time parameter may be either discrete or continuous, but since the latter is more important in economic applications, we concentrate on it.

Suppose that on a basic probability space there is defined a *history* $\mathscr{H} = (\mathscr{H}_t)_{t\geq 0}$ representing observable events as a function of time. For each t, $(\mathscr{H}_t)$ is the σ-algebra comprising events determined by observations over the interval $[0, t]$, so that $(\mathscr{H}_s) \subseteq (\mathscr{H}_t)$ when $s \leq t$. Then a stochastic process $M = (M_t)_{t\geq 0}$ is a *martingale* with respect to this history if

(a) For each t, M_t is $\mathscr{H}_t$ measurable (i.e., the state of the process at t is observable over $[0, t]$);
(b) $E[|M_t|] < \infty$ for each t;
(c) The 'martingale property' holds: whenever $s \leq t$,

$$E[M_t|\,\mathscr{H}_s] = M_s. \tag{1}$$

When no history is specified, it is usually understood that $\mathscr{H}_t = \sigma(M_s; s \leq t)$. One specific consequence is that $E[M_t] = E[M_0]$ for each t, so that a martingale is constant in the mean.

Written as

$$E[M_t - M_s|\ \mathscr{H}_s] = 0, s \leq t,$$

the martingale property implies that the optimal (in the sense of minimum mean squared error, or MMSE) predictor of a future increment of a martingale is zero. Thus, a martingale is indeed a mathematical idealization of a fair game. In some ways this property is clearest in differential form: assuming that the differential dM_t, which always extends *forward* in time from t, can be defined, then M is a martingale provided that

$$E[dM_t|\,\mathscr{H}_t] = 0 \tag{2}$$

for each t. Thus, a martingale can be interpreted as a 'noise' process, in which the MMSE prediction of the differential dM_t is simply zero; in many applications this interpretation becomes quite literal. Martingales are also analogous to the residuals in a regression problem, where what remains unexplained by the model should reduce, ideally, to chance variation.

One can also define *supermartingales*, for which (1) becomes

$$E[M_t|\,\mathscr{H}_t] \leq M_s, \tag{3}$$

and *submartingales*, in which the sense of the inequality in (3) is reversed. A supermartingale represents a less-than-fair game.

All martingales are in some sense convex combinations of (generalizations of) two key examples, namely the Wiener and Poisson processes. If (W_t) is a Wiener process (Brownian motion), then the processes W_t and $W_t^2 - t$ are both martingales; in fact, these properties characterize the Wiener process. In discrete time, martingales generalize sums of independent, mean zero random variables; the Wiener process, which has independent

and stationary increments, is a continuous time counterpart of these partial sum processes.

If (N_t) is a point process (or counting process), with N_t the number of events occurring in $[0, t]$, then under quite general assumptions there exists a nonnegative, predictable (a technical term, which in practice means left-continuous) random process (λ_t), the stochastic intensity of N, such that the process $M_t = N_t - \int_0^t \lambda_s ds$ is a martingale. Since $\lambda_t dt = E[dN_t | \mathscr{H}_t]$, M represents the new information realized as a function of time and, because of this and applications in statistics and state estimation, is known as the innovation martingale. For a Poisson process, which like the Wiener process has independent and stationary increments, the stochastic intensity is deterministic and equal to the rate of the process.

Square integrable martingales are especially important. A martingale M is *square integrable* if $\sup_t E\left[M_t^2\right] < \infty$, and in this case there exists a predictable process $\langle M \rangle$, the *predictable variation* of M, such that $M_t^2 - \langle M \rangle_t$ is a martingale. That the predictable variation is incrementally a conditional variance is confirmed by the differential relationship

$$\begin{aligned} d\langle M \rangle_t &= E[(dM_t - E dM_t | \mathscr{H}_t])^2 | \mathscr{H}_t] \\ &= E\left[(dM_t)^2 | \mathscr{H}_t\right]. \end{aligned}$$

Here the second equality holds because M is a martingale.

For the Wiener process $\langle W \rangle_t \equiv t$ in particular, the predictable variation is deterministic, a property characteristic of processes with independent increments. For a point process N with stochastic intensity λ, the predictable variation of the innovation martingale $dM_t = dN_t - \lambda_t dt$ is given by $d\langle M \rangle_t = \lambda_t dt$, which implies that a point process is locally and conditionally Poisson, in the sense that the incremental conditional mean and variance coincide.

Existence of the predictable variation is proved via the Doob-Meyer decomposition theorem, a cornerstone of the theory. The principal theoretical results pertaining to martingales fall into three classes: inequalities, convergence theorems and optimal sampling theorems.

So-called maximal inequalities, which provide upper bounds for probabilities of the form $P\{sup_{s \le t} |M_s| > c\}$, are not only of inherent interest, but also the key tools for proving convergence theorems. Moreover, these inequalities form the basis of a profound connection between martingales and classical mathematical analysis.

Under various assumptions, given a martingale M there exists a random variable M_∞ such that $M_t \to M_\infty$ almost surely as $t \to \infty$. Convergence obtains both almost surely and in L^1 if M is uniformly integrable, and in this case $M_t = E[M_\infty | \mathscr{H}_t]$ for each t. Not all martingles converge, however; those that fail to converge include, for example, the Wiener process and most innovation martingales.

Optional sampling theorems require the further concept of a stopping time. A random time T (a random variable with values in $[0, \infty]$, interpreted as the time at which some event occurs – with $T = \infty$ corresponding to its not occurring) is a *stopping time* of the history $\mathscr{H}$ if $\{T \le t\} \epsilon \mathscr{H}_t$ for each t. Intuitively, whether a stopping time has occurred by t can be determined from observations over $[0, t]$, and does not require prescient knowledge of the future. The rule by which a gambler quits a game must be a stopping time. Associated with a stopping time T is a σ-algebra $\mathscr{H}_t$ representing events determined by observations over the random time interval $[0, T]$ in the same way that for deterministic t, $\mathscr{H}_t$ corresponds to the interval $[0, t]$.

Martingale property extends from deterministic times to stopping times, and imply in particular that an unfair game cannot be made fair by means of a stopping time. More precisely, if M is a martingale and S and T are stopping times with $S \le T$, then under broad – albeit not universal – conditions,

$$E[M_T | \mathscr{H}_s] \le M_s. \tag{4}$$

With $S = 0$ in (4), taking expectations yields $E[M_T] = E[M_0]$. The corresponding result for supermartingales,

$$E[M_T|\mathscr{H}_s] \leq M_s, \tag{5}$$

demonstrates that an unfair game cannot be made fair via a stopping time, and dooms gambling systems without infinite resources to eventual failure.

Significant applications of martingales include mathematical statistics (likelihood ratio processes are martingales), queueing theory, filtering and prediction (for example, in signal processing) and economics.

A common feature of these applications is that they involve a random system 'driven' by a martingale in precisely the same manner that a dynamical system is driven by a forcing function. Given a (square integrable) martingale M and a predictable process C fulfilling integrability restrictions, the stochastic integral process

$$(C * M)_t = \int_0^t C_s dM_s$$

is itself a martingale, for which M acts as driving term. (Since M may change state discontinuously, whether endpoints are included in the interval of integration must be specified; in this case, the integral is over the closed interval $[0, t]$.) Construction of stochastic integrals is a difficult, subtle problem: none of the conventional definitions can be applied pathwise (typically the sample paths of M are not of bounded variation), and instead one must employ sophisticated probability theory. The predictable variations satisfy

$$d\langle C * M\rangle_t = C_t^2 d\langle M\rangle_t.$$

Economic applications include, e.g., models of securities prices.

In applications, the inclusion of a 'dt'-integral is often desirable or necessary, leading to *semimartingales,* which are random processes Z of the form

$$Z_t = \int_0^t A_s ds + \int_0^t C_s dM_s, \tag{6}$$

where M is a martingale, C is a predictable process and A fulfills a technical property known as progressive measurability. (Integrability conditions must be satisfied as well.) The differential version of (6) is

$$dZ_t = A_t dt + C_t dM_t. \tag{7}$$

If the processes A and C, rather than specified exogenously, are functionals of Z, then (7) becomes a stochastic differential equation

$$dZ_t = \mu(Z_t)dt + \sigma(Z_t)dM_t, \tag{8}$$

or, more generally,

$$dZ_t = \mu(Z_s; s \leq t)dt + \sigma(Z_s; s \leq t)dM_t. \tag{9}$$

These equations can be solved – however, not using pathwise methods – under a variety of assumptions, but essentially only when the driving term is a martingale. For example, if the martingale is the Wiener process, solutions to (8) and (9) are known as diffusions and Itō processes, respectively, and the resultant theory as the Itō calculus, after its principal inventor, K. Itō. Alternatively, if M is the innovation martingale associated with a point process N then, inter alia, solutions to (8) can be used to construct recursive methods for filtering to extract signals from noise.

Bibliography

Brémaud, P. 1981. *Point processes and queues: Martingale dynamics*. Berlin: Springer-Verlag.

Hall, P., and C.C. Heyde. 1980. *Martingale limit theory and its applications*. New York: Academic Press.

Kallianpur, G. 1980. *Stochastic filtering theory.* New York: Springer-Verlag.

Karr, A.F. 1986. *Point processes and their statistical inference*. New York: Marcel Dekker.

Lipster, R.S., and A.N. Shiryayev. 1978. *Statistics of random processes, I and II*. Berlin: Springer-Verlag.

Metivier, M., and J. Pellaumail. 1980. *Stochastic integration*. New York: Academic Press.

Shiryayev, A.N. 1981. Martingales: Recent developments, results and applications. *International Statistical Review* 49: 199–233.

Marx, Karl Heinrich (1818–1883)

Ernest Mandel

Abstract

This article summarizes the methodology and economics of Karl Marx. After a brief account of his life, it deals with his historical materialism, and then his labour theory of value, his theories of rent, money, surplus value, and crises, his account of the laws of motion of the capitalism mode of production, and his and Engels's conception of the economy of post-capitalist societies.

Keywords

Accumulation of capital; Asiatic mode of production; Böhm-Bawerk, E. von; Bukharin, N.; Capitalist mode of production; Class; Commodity theory of money; Communism; Concentration (centralization) of capital; Constant and variable capital; Credit cycles; Economic crisis; Economic laws; Economic science; Engels, F.; Exploitation; Feudalism; Hilferding, R.; Historical materialism; Ideology; Invisible hand; Iron law of wages; Labour power; Labour theory of value; Land rent; Lassalle, F.; Laws of motion of capitalism; Lenin, V.; Market value; Marx, K.; Mobility of capital; Modes of production; Monopoly; Organic composition of capital; Overproduction; Paper money; Post-capitalist society; Prices of production; Proletariat; Rate of profit; Rent; Reserve army of labour; Ricardo, D.; Social polarization; Socialism; Sraffa, P.; Surplus value; Technological change; Trade unions; Transformation problem; Underconsumptionism

JEL Classifications

B31

Karl Marx was born on 5 May 1818, the son of the lawyer Heinrich Marx and Henriette Pressburg. His father was descended from an old family of Jewish rabbis, but was himself a liberal admirer of the Enlightenment and not religious. He converted to Protestantism a few years before Karl was born to escape restrictions still imposed upon Jews in Prussia. His mother was of Dutch-Jewish origin.

Life and Work

Karl Marx studied at the *Friedrich-Wilhelm Gymnasium* in Trier, and at the universities of Bonn and Berlin. His doctoral thesis, *Differenz der demokritischen und epikurischen Naturphilosophie*, was accepted at the University of Jena on 15 April 1841. In 1843 he married Jenny von Westphalen, daughter of Baron von Westphalen, a high Prussian government official.

Marx's university studies covered many fields, but centred around philosophy and religion. He frequented the circle of the more radical followers of the great philosopher Hegel, befriended one of their main representatives, Bruno Bauer, and was especially influenced by the publication in 1841 of Ludwig Feuerbach's *Das Wesen des Christentums* (The Nature of Christianity). He had intended to teach philosophy at the university, but that quickly proved to be unrealistic. He then turned towards journalism, both to propagandize his ideas and to gain a livelihood. He became editor of the *Rheinische Zeitung,* a liberal newspaper of Cologne, in May 1942. His interest turned more and more to political and social questions, which he treated in an increasingly radical way. The paper was banned by the Prussian authorities a year later.

Karl Marx then planned to publish a magazine called *Deutsch-Französische Jahrbücher* in Paris, in order to escape Prussian censorship and to be more closely linked and identified with the real struggles for political and social emancipation which, at that time, were centred around France. He emigrated to Paris with his wife and met there his lifelong friend Friedrich Engels.

Marx had become critical of Hegel's philosophical political system, a criticism which would lead to his first major work, Zur Kritik

des *Hegelschen Rechtsphilosophie* (A Critique of Hegel's Philosophy of Right). Intensively studying history and political economy during his stay in Paris, he became strongly influenced by socialist and working-class circles in the French capital. With his 'Paris Manuscripts' (*Oekonomisch-philosophische Manuskripte,* 1844), he definitely became a communist, i.e. a proponent of collective ownership of the means of production.

He was expelled from France at the beginning of 1845 through pressure from the Prussian embassy and migrated to Brussels. His definite turn towards historical materialism (see below) would occur with his manuscript *Die Deutsche Ideologie* (1845–6) culminating in the eleven *Theses on Feuerbach,* written together with Engels but never published during his lifetime.

This led also to a polemical break with the most influential French socialist of that period, Proudhon, expressed in the only book Marx would write in French, *Misère de la Philosophie* (1846).

Simultaneously he became more and more involved in practical socialist politics, and started to work with the Communist League, which asked Engels and himself to draft their declaration of principle. This is the origin of the *Communist Manifesto* (1848), *Manifest der Kommunistischen Partei* (1848).

As soon as the revolution of 1848 broke out, he was in turn expelled from Belgium and went first to France, then, from April 1848 on, to Cologne. His political activity during the German revolution of 1848 centred around the publication of the daily paper *Neue Rheinische Zeitung*, which enjoyed wide popular support. After the victory of the Prussian counter-revolution, the paper was banned in May 1849 and Marx was expelled from Prussia. He never succeeded in recovering his citizenship. Marx emigrated to London, where he would stay, with short interruptions, till the end of his life. For fifteen years, his time would be mainly taken up with economic studies, which would lead to the publication first of *Zur Kritik der Politischen Oekonomie* (1859) and later of *Das Kapital,* Vol. I (1867). He spent long hours at the British Museum, studying the writings of all the major economists, as well as the government Blue Books, Hansard and many other contemporary sources on social and economic conditions in Britain and the world. His reading also covered technology, ethnology and anthropology, besides political economy and economic history; many notebooks were filled with excerpts from the books he read.

But while the activity was mainly studious, he never completely abandoned practical politics. He first hoped that the Communist League would be kept alive, thanks to a revival of revolution. When this did not occur, he progressively dropped out of emigré politics, but not without writing a scathing indictment of French counter-revolution in *Der 18. Brumaire des Louis Bonaparte* (1852), which was in a certain sense the balance sheet of his political activity and an analysis of the 1848–52 cycle of revolution and counter-revolution. He would befriend British trade-union leaders and gradually attempt to draw them towards international working class interests and politics. These efforts culminated in the creation of the International Working Men's Association (1864) – the so-called First International – in which Marx and Engels would play a leading role, politically as well as organizationally.

It was not only his political interest and revolutionary passion that prevented Marx from becoming an economist pure and simple. It was also the pressure of material necessity. Contrary to his hope, he never succeeded in earning enough money from his scientific writings to sustain himself and his growing family. He had to turn to journalism to make a living. He had initial, be it modest, success in this field, when he became European correspondent of the *New York Daily Tribune* in the summer of 1851. But he never had a regular income from that collaboration, and it ended after ten years.

So the years of his London exile were mainly years of great material deprivation and moral suffering. Marx suffered greatly from the fact that he could not provide a minimum of normal living conditions for his wife and children, whom he loved deeply. Bad lodgings in cholera-stricken Soho, insufficient food and medical care, led to a chronic deterioration of his wife's and his own health and to the death of several of their children; that of his oldest son Edgar in 1855 struck him an

especially heavy blow. Of his seven children, only three daughters survived, Jenny, Laura and Eleanor (Tussy). All three were very gifted and would play a significant role in the international labour movement, Eleanor in Britain, Jenny and Laura in France (where they married the socialist leaders Longuet and Lafargue).

During this long period of material misery, Marx survived thanks to the financial and moral support of his friend Friedrich Engels, whose devotion to him stands as an exceptional example of friendship in the history of science and politics. Things started to improve when Marx came into his mother's inheritance; when the first independent working-class parties (followers of Lassalle on the one hand, of Marx and Engels on the other) developed in Germany, creating a broader market for his writings; when the IWMA became influential in several European countries, and when Engels' financial conditions improved to the point where he would sustain the Marx family on a more regular basis.

The period 1865–71 was one in which Marx's concentration on economic studies and on the drafting of *Das Kapital* was interrupted more and more by current political commitments to the IWMA, culminating in his impassioned defence of the Paris Commune (*Der Bürgerkrieg in Frankreich [The Civil War in France]* 1871). But the satisfaction of being able to participate a second time in a real revolution – be it only vicariously – was troubled by the deep divisions inside the IMWA, which led to the split with the anarchists grouped around Michael Bakunin.

Marx did not succeed in finishing a final version of *Das Kapital* vols II and III, which were published posthumously, after extensive editing, by Engels. It remains controversial whether he intended to add two more volumes to these, according to an initial plan. More than 25 years after the death of Marx, Karl Kautsky edited what is often called vol. IV of *Das Kapital,* his extensive critique of other economists: *Theorien über den Mehrwert (Theories of Surplus Value).*

Marx's final years were increasingly marked by bad health, in spite of slightly improved living conditions. Bad health was probably the main reason why the final version of vols II and III of *Capital* could not be finished. Although he wrote a strong critique of the Programme which was adopted by the unification congress (1878) of German social democracy *(Kritik Des Gothaer Program),* he was heartened by the creation of that united working-class party in his native land, by the spread of socialist organizations throughout Europe, and by the growing influence of his ideas in the socialist movement. His wife fell ill in 1880 and died the next year. This came as a deadly blow to Karl Marx, who did not survive her for long. He himself died in London on 14 March 1883.

Historical Materialism

Outside his specific economic theories, Marx's main contribution to the social sciences has been his theory of historical materialism. Its starting point is anthropological. Human beings cannot survive without social organization. Social organization is based upon social labour and social communication. Social labour always occurs within a given framework of specific, historically determined, social relations of production. These social relations of production determine in the last analysis all other social relations, including those of social communication. It is social existence which determines social consciousness and not the other way around.

Historical materialism posits that relations of production which become stabilized and reproduce themselves are structures which can no longer be changed gradually, piecemeal. They are modes of production. To use Hegel's dialectical language, which was largely adopted (and adapted) by Marx: they can only change qualitatively through a complete social upheaval, a social revolution or counter-revolution. Quantitative changes can occur within modes of production, but they do not modify the basic structure. In each mode of production, a given set of relations of production constitutes the basis (infrastructure) on which is erected a complex superstructure, encompassing the state and the law (except in classless society), ideology, religion, philosophy, the arts, morality etc.

Relations of production are the sum total of social relations which human beings establish among themselves in the production of their material lives. They are therefore not limited to what actually happens at the point of production. Humankind could not survive, i.e. produce, if there did not exist specific forms of circulation of goods, e.g. between producing units (circulation of tools and raw materials) and between producing units and consumers. A priori allocation of goods determines other relations of production than does allocation of goods through the market. Partial commodity production (what Marx calls 'simple commodity production' or 'petty commodity production' – 'einfache Warenproduktion') also implies other relations of production than does generalized commodity production.

Except in the case of classless societies, modes of production, centred around prevailing relations of production, are embodied in specific class relations which, in the last analysis, overdetermine relations between individuals.

Historical materialism does not deny the individual's free will, his attempts to make choices concerning his existence according to his individual passions, his interests as he understands them, his convictions, his moral options etc. What historical materialism does state is: (1) that these choices are strongly predetermined by the social framework (education, prevailing ideology and moral 'values', variants of behaviour limited by material conditions etc); (2) that the outcome of the collision of millions of different passions, interests and options is essentially a phenomenon of social logic and not of individual psychology. Here, class interests are predominant.

There is no example in history of a ruling class not trying to defend its class rule, or of an exploited class not trying to limit (and occasionally eliminate) the exploitation it suffers. So outside classless society, the class struggle is a permanent feature of human society. In fact, one of the key theses of historical materialism is that 'the history of humankind is the history of class struggles' (Marx, *Communist Manifesto,* 1848).

The immediate object of class struggle is economic and material. It is a struggle for the division of the social product between the direct producers (the productive, exploited class) and those who appropriate what Marx calls the social surplus product, the residuum of the social product once the producers and their offspring are fed (in the large sense of the word; i.e. the sum total of the consumer goods consumed by that class) and the initial stock of tools and raw materials is reproduced (including the restoration of initial fertility of the soil). The ruling class functions as ruling class essentially through the appropriation of the social surplus product. By getting possession of the social surplus product, it acquires the means to foster and maintain most of the superstructural activities mentioned above; and by doing so, it can largely determine their function – to maintain and reproduce the given social structure, the given mode of production – and their contents.

We say 'largely determine' and not 'completely determine'. First, there is an 'immanent dialectical', i.e. an autonomous movement, of each specific superstructural sphere of activity. Each generation of scientists, artists, philosophers, theologists, lawyers and politicians finds a given *corpus* of ideas, forms, rules, techniques, ways of thinking, to which it is initiated through education and current practice, etc. It is not forced to simply continue and reproduce these elements. It can transform them, modify them, change their interconnections, even negate them. Again: historical materialism does not deny that there is a specific history of science, a history of art, a history of philosophy, a history of political and moral ideas, a history of religion etc., which all follow their own logic. It tries to *explain* why a certain number of scientific, artistic, philosophical, ideological, juridical changes or even revolutions occur at a given time and in given countries, quite different from other ones which occurred some centuries earlier elsewhere. The nexus of these 'revolutions' with given historical periods is a nexus of class interests.

Second, each social formation (i.e. a given country in a given epoch) while being characterized by predominant relations of production (i.e. a given mode or production at a certain phase of its development) includes different relations of production which are largely remnants of the past, but

also sometimes nuclei of future modes of production. Thus there exists not only the ruling class and the exploited class characteristic of that prevailing mode of production (capitalists and wage earners under capitalism). There also exist remnants of social classes which were predominant when other relations of production prevailed and which, while having lost their hegemony, still manage to survive in the interstices of the new society. This is for example the case with petty commodity producers (peasants, handicraftsmen, small merchants), semifeudal landowners, and even slave-owners, in many already predominantly capitalist social formations throughout the 19th and part of the 20th centuries. Each of these social classes has its own ideology, its own religious and moral values, which are intertwined with the ideology of the hegemonic ruling class, without becoming completely absorbed by that.

Third, even after a given ruling class (e.g. the feudal or semi-feudal nobility) has disappeared as a ruling class, its ideology can survive through sheer force of social inertia and routine (custom). The survival of traditional *ancien régime* catholic ideology in France during a large part of the 19th century, in spite of the sweeping social, political and ideological changes ushered in by the French revolution, is an illustration of that rule.

Finally, Marx's statement that the ruling ideology of each epoch is the ideology of the ruling class – another basic tenet of historical materialism – does not express more than it actually says. It implies that other ideologies can exist side by side with that ruling ideology without being hegemonic. To cite the most important of these occurrences: exploited and (or) oppressed social classes can develop their own ideology, which will start to challenge the prevailing hegemonic one. In fact, an ideological class struggle accompanies and sometimes even precedes the political class struggle properly speaking. Religious and philosophical struggles preceding the classical bourgeois revolutions; the first socialist critiques of bourgeois society preceding the constitution of the first working-class parties and revolutions, are examples of that type.

The class struggle has been up to now the great motor of history. Human beings make their own history. No mode of production can be replaced by another one without deliberate actions by large social forces, i.e., without social revolutions (or counter-revolutions). Whether these revolutions or counter-revolutions actually lead to the long-term implementation of deliberate projects of social reorganization is another matter altogether. Very often, their outcome is to a large extent different from the intention of the main actors.

Human beings act consciously, but they can act with false consciousness. They do not necessarily understand why they want to realize certain social and (or) political plans, why they want to maintain or to change economic or juridical institutions; and especially, they rarely understand in a scientific sense the laws of social change, the material and social preconditions for successfully conserving or changing such institutions. Indeed, Marx claims that only with the discovery of the main tenets of historical materialism have we made a significant step forward towards understanding these laws, without being able to predict 'all' future developments of society.

Social change, social revolutions and counter-revolutions are furthermore occurring within determined material constraints. The level of development of the productive forces – essentially tools and human skills, including their effects upon the fertility of the soil – limits the possibilities of institutional change. Slave labour has shown itself to be largely incompatible with the factory system based upon contemporary machines. Socialism would not be durably built upon the basis of the wooden plough and the potter's wheel. A social revolution generally widens the scope for the development of the productive forces and leads to social progress in most fields of human activity in a momentous way. Likewise, an epoch of deep social crisis is ushered in when there is a growing conflict between the prevailing mode of production (i.e. the existing social order) on the one hand, and the further development of the productive forces on the other. Such a social crisis will then manifest itself on all major fields and social activity: politics, ideology, morals and law, as well as in the realm of the economic life properly speaking.

Historical materialism thereby provides a measuring stick for human progress: the growth of the productive forces, measurable through the growth of the average productivity of labour, and the number, longevity and skill of the human species. This measuring stick in no way abstracts from the natural preconditions for human survival and human growth (in the broadest sense of the concept). Nor does it abstract from the conditional and partial character of such progress, in terms of social organization and individual alienation.

In the last analysis, the division of society into antagonistic social classes reflects, from the point of view of historical materialism, an inevitable limitation of human freedom. For Marx and Engels, the real measuring rod of human freedom, i.e. of human wealth, is not 'productive labour'; this only creates the material pre-condition for that freedom. The real measuring rod is leisure time, not in the sense of 'time for doing nothing' but in the sense of time freed from the iron necessity to produce and reproduce material livelihood, and therefore disposable for all-round and free development of the individual talents, wishes, capacities, potentialities, of each human being.

As long as society is too poor, as long as goods and services satisfying basic needs are too scarce, only part of society can be freed from the necessity to devote most of its life to 'work for a livelihood' (i.e. of forced labour, in the anthropological/sociological sense of the word, that is in relation to desires, aspirations and talents, not to a juridical status of bonded labour). That is essentially what represents the freedom of the ruling classes and their hangers-on, who are 'being paid to think', to create, to invent, to administer, because they have become free from the obligation to bake their own bread, weave their own clothes and build their own houses.

Once the productive forces are developed far enough to guarantee all human beings satisfaction of their basic needs by 'productive labour' limited to a minor fraction of lifetime (the half work-day or less), then the material need of the division of society in classes disappears. Then, there remains no objective basis for part of society to monopolize administration, access to information, knowledge, intellectual labour. For that reason, historical materialism explains both the reasons why class societies and class struggles arose in history, and why they will disappear in the future in a classless society of democratically self-administering associated producers.

Historical materialism therefore contains an attempt at explaining the origin, the functions, and the future withering away of the state as a specific institution, as well as an attempt to explain politics and political activity in general, as an expression of social conflicts centred around different social interest (mainly, but not only, those of different social classes; important fractions of classes, as well as non-class social groupings, also come into play).

For Marx and Engels, the state is not existent with human society as such, or with 'organized society' or even with 'civilized society' in the abstract; neither is it the result of any voluntarily concluded 'social contract' between individuals. That state is the sum total of apparatuses, i.e. special groups of people separate and apart from the rest (majority) of society, that appropriate to themselves functions of a repressive or integrative nature which were initially exercised by all citizens. This process of alienation occurs in conjunction with the emergence of social classes. The state is an instrument for fostering, conserving and reproducing a given class structure, and not a neutral arbiter between antagonistic class interests.

The emergence of a classless society is therefore closely intertwined, for adherents to historical materialism, with the process of withering away of the state, i. e. of gradual devolution to the whole of society (self-management, self-administration) of all specific functions today exercised by special apparatuses, i.e. of the dissolution of these apparatuses. Marx and Engels visualized the dictatorship of the proletariat, the last form of the state and of political class rule, as an instrument for assuring the transition from class society to classless society. It should itself be a state of a special kind, organizing its own gradual disappearance.

We said above that, from the point of view of historical materialism, the immediate object of class struggle is the division of the social product

between different social classes. Even the political class struggle in the final analysis serves that main purpose; but it also covers a much broader field of social conflicts. As all state activities have some bearing upon the relative stability or instability of a given social formation, and the class rule to which it is submitted, the class struggle can extend to all fields of politics, from foreign policy to educational problems and religious conflicts. This has of course to be proven through painstaking analysis, and not proclaimed as an axiom or a revealed truth. When conducted successfully, such exercises in class analysis and class definition of political, social and even literary struggles become impressive works of historical explanation, as for example Marx's *Class Struggles in France 1848–50*, Engels' *The German Peasant War*, Franz Mehring's *Die Lessing-Legende*, Trotsky's *History of the Russian Revolution*, etc.

Marx's Economic Theory: General Approach and Influence

A general appraisal of Marx's method of economic analysis is called for prior to an outline of his main economic theories (theses and hypotheses).

Marx is distinct from most important economists of the 19th and 20th centuries in that he does not consider himself at all an 'economist' pure and simple.

The idea that 'economic science' as a special science completely separate from sociology, history, anthropology etc. cannot exist, underlies most of his economic analysis. Indeed, historical materialism is an attempt at unifying all social sciences, if not all sciences about humankind, into a single 'science of society'.

For sure, within the framework of this general 'science of society', economic phenomena could and should be submitted to analysis as specific phenomena. So economic theory, economical science, have a definite autonomy after all; but is only a partial and relative one.

Probably the best formula for characterizing Marx's economic theory would be to call it an endeavour to explain the social economy. This would be true in a double sense. For Marx, there are no eternal economic laws, valid in every epoch of human prehistory and history. Each mode of production has its own specific economic laws, which lose their relevance once the general social framework has fundamentally changed. For Marx likewise, there are no economic laws separate and apart from specific relations between human beings, in the primary (but not only, as already summarized) social relations of production. All attempts to reduce economic problems to purely material, objective ones, to relations between things, or between things and human beings, would be considered by Marx as manifestations of mystification, of false consciousness, expressing itself through the attempted reification of human relations. Behind relations between things, economic science should try to discover the specific relations between human beings which they hide. Real economic science has therefore also a demystifying function compared to vulgar 'economics', which takes a certain number of 'things' for granted without asking the question: Are they really only what they appear to be? From where do they originate? What explains these appearances? What lies behind them? Where do they lead? How could they (will they) disappear? *Problemblindheit,* the refusal to see that facts are generally more problematic than they appear at first sight, is certainly not a reproach one could address to Marx's economic thought.

Marx's economic analysis is therefore characterized by a strong ground current of *historical relativism,* with a strong recourse to the genetical and evolutionary method of thinking (that is why the parallel with Darwin has often been made, sometimes in an excessive way). The formula 'genetic structuralism' has also been used in relation to Marx's general approach to economic analysis. Be that as it may, one could state that Marx's economic theory is essentially geared to the discovery of specific 'laws of motion' for successive modes of production. While his theoretical effort has been mainly centred around the discovery of these laws of motion for capitalist society, his work contains indications of such laws – different ones, to be sure – for precapitalist and post-capitalist social formations too.

The main link between Marx's sociology and anthropology on the one hand, and his economic analysis on the other, lies in the key role of social labour as the basic anthropological feature underlying all forms of social organization. Social labour can be organized in quite different forms, thereby giving rise to quite different economic phenomena ('facts'). Basically different forms of social labour organization lead to basically different sets of economic institutions and dynamics, following basically different logics (obeying basically different 'laws of motion').

All human societies must assure the satisfaction of a certain number of basic needs, in order to survive and reproduce themselves. This leads to the necessity of establishing some sort of equilibrium between socially recognized needs, i.e. current consumption and current production. But this abstract banality does not tell us anything about the concrete way in which social labour is organized in order to achieve that goal.

Society can recognize all individual labour as *immediately social labour.* Indeed, it does so in innumerable primitive tribal and village communities, as it does in the contemporary *kibbutz.* Directly social labour can be organized in a despotic or in a democratic way, through custom and superstition as well as through an attempt at applying advanced science to economic organization; but it will always be immediately recognized social labour, in as much as it is based upon *a prior* assignment of the producers to their specific work (again: irrespective of the form this assignation takes, whether it is voluntary or compulsory, despotic or simply through custom etc.).

But when social decision-taking about work assignation (and resource allocation closely tied to it) is fragmented into different units operating independently from each other – as a result of private control (property) of the means of production, in the economic and not necessarily the juridical sense of the word – then social labour in turn is fragmented into private labours which are not automatically recognized as socially necessary ones (whose expenditure is not automatically compensated by society). Then the private producers have to exchange parts or all of their products in order to satisfy some or all of their basic needs. Then these products become commodities. The economy becomes a (partial or generalized) market economy. Only by measuring the results of the sale of his products can the producer (or owner) ascertain what part of his private labour expenditure has been recognized (compensated) as social labour, and what part has not.

Even if we operate with such simple analytical tools as 'directly social labour', 'private labour', 'socially recognized social labour', we have to make quite an effort at abstracting from immediately apparent phenomena in order to understand their relevance for economic analysis. This is true for all scientific analysis, in natural as well as in social sciences. Marx's economic analysis, as presented in his main books, has not been extremely popular reading; but then, there are not yet so many scientists in these circumstances. This has nothing to do with any innate obscurity of the author, but rather with the nature of scientific analysis as such.

The relatively limited number of readers of Marx's economic writings (the first English paperback edition of *Das Kapital* appeared only in 1974!) is clearly tied to Marx's scientific rigour, his effort at a systematic and all-sided analysis of the phenomena of the capitalist economy.

But while his economic analysis lacked popularity, his political and historical projections became more and more influential. With the rise of independent working-class mass parties, an increasing number of these proclaimed themselves as being guided or influenced by Marx, at least in the epoch of the Second and the Third Internationals, roughly the half century from 1890 till 1940. Beginning with the Russian revolution of 1917, a growing number of governments and of states claimed to base their policies and constitutions on concepts developed by Marx. (Whether this was legitimate or not is another question.) But the fact itself testifies to Marx's great influence on contemporary social and political developments, evolutionary and revolutionary alike.

Likewise, his diffused influence on social science, including academic economic theory, goes far beyond general acceptance or even substantial knowledge of his main writings. Some key ideas of historical materialism and of economic analysis

which permeate his work – e.g. that economic interests to a large extent influence, if not determine, political struggles; that historic evolution is linked to important changes in material conditions; that economic crises ('the business cycle') are unavoidable under conditions of capitalist market economy – have become near-platitudes. It is sufficient to notice how major economists and historians strongly denied their validity throughout the 19th century and at least until the 1920s, to understand how deep has been Marx's influence on contemporary social science in general.

Marx's Labour Theory of Value

As an economist, Marx is generally situated in the continuity of the great classical school of Adam Smith and Ricardo. He obviously owes a lot to Ricardo, and conducts a current dialogue with that master in most of his mature economic writings.

Marx inherited the labour theory of value from the classical school. Here the continuity is even more pronounced; but there is also a radical break. For Ricardo, labour is essentially a numeraire, which enables a common computation of labour and capital as basic elements of production costs. For Marx, labour is value. Value is nothing but that fragment of the total labour potential existing in a given society in a certain period (e.g. a year or a month) which is used for the output of a given commodity, at the average social productivity of labour existing then and there, divided by the total number of these commodities produced, and expressed in hours (or minutes), days, weeks, months of labour.

Value is therefore essentially a social, objective and historically relative category. It is social because it is determined by the overall result of the fluctuating efforts of each individual producer (under capitalism: of each individual firm or factory). It is objective because it is given, once the production of a given commodity is finished and is thus independent from personal (or collective) valuations of customers on the market place; and it is historically relative because it changes with each important change (progress or regression) of the average productivity of labour in a given branch of output, including in agriculture and transportation.

This does not imply that Marx's concept of value is in any way completely detached from consumption. It only means that the feedback of consumers' behaviour and wishes upon value is always mediated through changes in allocation of labour inputs in production, labour seen as subdivided into living labour and dead (dated) labour, i.e. tools and raw materials. The market emits signals to which the producing units react. Value changes after these reactions, not before them. Market price changes can of course occur prior to changes in value. In fact, changes in market prices are among the key signals which can lead to changes in labour allocation between different branches of production, i.e. to changes in labour quantities necessary to produce given commodities. But then, for Marx, values determine prices only basically and in the medium-term sense of the word. This determination only appears clearly as an explication of *medium and long-term price movements.* In the shorter run, prices fluctuate around values as axes. Marx never intended to negate the operation of market laws, of the law of supply and demand, in determining these short-term fluctuations.

The 'law of value' is but Marx's version of Adam Smith's 'invisible hand'. In a society dominated by private labour, private producers and private ownership of productive inputs, it is this 'law of value', an objective economic law operating behind the backs of all people, all 'agents' involved in production and consumption, which, in the final analysis, regulates the economy, determines what is produced and how it is produced (and therefore also what can be consumed). The 'law of value' regulates the exchange between commodities, according to the quantities of socially necessary abstract labour they embody (the quantity of such labour spent in their production). Through regulating the exchange between commodities, the 'law of value' also regulates, after some interval, the distribution of society's labour potential and of society's non-living productive resources between different branches of production. Again, the analogy with Smith's 'invisible hand' is striking.

Marx's critique of the 'invisible hand' concept does not dwell essentially on the analysis of how a market economy actually operates. It would above all insist that this operation is not eternal, not immanent in 'human nature', but created by specific historical circumstances, a product of a special way of social organization, and due to disappear at some stage of historical evolution as it appeared during a previous stage. And it would also stress that this 'invisible hand' leads neither to the maximum of economic growth nor to the optimum of human wellbeing for the greatest number of individuals, i.e. it would stress the heavy economic and social price humankind had to pay, and is still currently paying, for the undeniable progress the market economy produced at a given stage of historical evolution.

The formula 'quantities of abstract human labour' refers to labour seen strictly as a fraction of the total labour potential of a given society at a given time, say a labour potential of 2 billion hours a year (1 million potential producers supposedly capable of working each 2000 hours a year). It therefore implies making abstraction of the specific trade or occupation of a given male or female producer, the product of a day's work of a weaver not being worth less or more than that of a peasant, a miner, a housebuilder, a milliner or a seamstress. At the basis of that concept of 'abstract human labour' lies a social condition, a specific set of social relations of production, in which small independent producers are essentially equal. Without that equality, social division of labour, and therefore satisfaction of basic consumers' needs, would be seriously endangered under that specific organizational set-up of the economy. Such an equality between small commodity owners and producers is later transformed into an equality between owners of capital under the capitalist mode of production.

But the concept of homogeneity of productive human labour, underlying that of 'abstract human labour' as the essence of value, does not imply a negation of the difference between skilled and unskilled labour. Again: a negation of that difference would lead to breakdown of the necessary division of labour, as would any basic heterogeneity of labour inputs in different branches of output. It would then not pay to acquire skills: most of them would disappear. So Marx's labour theory of value, in an internally coherent way, leads to the conclusion that one hour of skilled labour represents more value than one hour of unskilled labour, say represents the equivalent of 1.5 hours of unskilled labour. The difference would result from the imputation of the labour it costs to acquire the given skill. While an unskilled labourer would have a labour potential of 120,000 hours during his adult life, a skilled labourer would only have a labour potential of 80,000 hours, 40,000 hours being used for acquiring, maintaining and developing his skill. Only if one hour of skilled labour embodies the same value of 1.5 hours of unskilled labour, will the equality of all 'economic agents' be maintained under these circumstances, i.e. will it 'pay' economically to acquire a skill.

Marx himself never extensively dwelled on this solution of the so-called *reduction problem.* This remains indeed one of the most obscure parts of his general economic theory. It has led to some, generally rather mild, controversy. Much more heat has been generated by another facet of Marx's labour theory of value, the so-called *transformation problem.* Indeed, from Böhm-Bawerk writing a century ago till the recent contributions of Sraffa (1960) and Steedman (1977), the way Marx dealt with the transformation of values into 'prices of production' in *Capital* Vol. III has been considered by many of his critics as the main problem of his 'system', including being a reason to reject the labour theory of value out of hand.

The problem arises out of the obvious modification in the functioning of a market economy when *capitalist* commodity production substitutes itself for *simple* commodity production. In simple commodity production, with generally stable technology and stable (or easily reproduceable) tools, living labour is the only variable of the quantity and subdivision of social production. The mobility of labour is the only dynamic factor in the economy. As Engels pointed out in his Addendum to *Capital* Vol. III (Marx, *g*, pp. 1034–7) in such an economy, commodities would be exchanged at prices which would be immediately proportional to values, to the labour inputs they embody.

But under the capitalist mode of production, this is no longer the case. Economic decision-taking is not in the hands of the direct producers. It is in the hands of the capitalist entrepreneurs in the wider sense of the word (bankers – distributors of credit – playing a key role in that decision-taking, besides entrepreneurs in the productive sector properly speaking). Investment decisions, i.e. decisions for creating, expanding, reducing or closing enterprises, determine economic life. It is the *mobility of capital* and not the mobility of labour which becomes the motive force of the economy. Mobility of labour becomes essentially an epiphenomenon of the mobility of capital.

Capitalist production is production for profit. Mobility of capital is determined by existing or expected profit differentials. Capital leaves branches (countries, regions) with lower profits (or profit expectations) and flows towards branches (countries, regions) with higher ones. These movements lead to an equalization of the rate of profit between different branches of production. But approximately equal returns on all invested capital (at least under conditions of prevailing 'free competition') coexist with unequal proportions of inputs of labour in these different branches. So there is a disparity between the direct value of a commodity and its 'price of production', that 'price of production' being defined by Marx as the sum of production costs (costs of fixed capital and raw materials plus wages) and the average rate of profit multiplied with the capital spent in the given production.

The so-called 'transformation problem' relates to the question of whether a relation can nevertheless be established between value and these ' prices of production', what is the degree of coherence (or incoherence) of the relation with the 'law of value' (the labour theory of value in general), and what is the correct quantitative way to express that relation, if it exists.

We shall leave aside here the last aspect of the problem, to which extensive analysis has recently been devoted (Mandel and Freeman 1984). From Marx's point of view, there is no incoherence between the formation of 'prices of production' and the labour theory of value. Nor is it true that he came upon that alleged difficulty when he started to prepare *Capital* III, i.e. to deal with capitalist competition, as several critics have argued (see e.g. Joan Robinson 1942). In fact, his solution of the transformation problem is already present in the *Grundrisse* (Marx, *d*), before he even started to draft *Capital* Vol. I.

The sum total of value produced in a given country during a given span of time (e.g. one year) is determined by the sum total of labour-inputs. Competition and movements of capital cannot change that quantity. The sum total of values equals the sum total of 'prices of production'. The only effect of capital competition and capital mobility is to *redistribute* that given sum – and this through a redistribution of surplus value (see below) – between different capitals, to the benefit of some and at the expense of others.

Now this redistribution does not occur in a haphazard or arbitrary way. Essentially value (surplus-value) is transferred from technically less advanced branches to technologically more advanced branches. And here the concept of 'quantities of socially necessary labour' comes into its own, under the conditions of constant revolutions of productive technology that characterize the capitalist mode of production. Branches with lower than average technology (organic composition of capital, see below) can be considered as wasting socially necessary labour. Part of the labour spent in production in their realm is therefore not compensated by society. Branches with higher than average technology (organic composition of capital) can be considered to be economizing social labour; their labour inputs can therefore be considered as more intensive than average, embodying more value. In this way, the transfer of value (surplus-value) between different branches, far from being in contradiction with the law of value, is precisely the way it operates and should operate under conditions of 'capitalist equality', given the pressure of rapid technological change.

As to the logical inconsistency often supposedly to be found in Marx's method of solving the 'transformation problem' – first advanced by von Bortkiewicz (1907) – it is based upon a misunderstanding in our opinion. It is alleged that in his 'transformation schemas' (or tables) (Marx, *g*, pp. 255–6) Marx calculates inputs in 'values'

and outputs in 'prices of production', thereby omitting the feedback effect of the latter on the former. But that feedback effect is unrealistic and unnecessary, once one recognizes that inputs are essentially data. Movements of capital posterior to the purchase of machinery or raw materials, including ups and downs of prices of finished products produced with these raw materials, cannot lead to a change in prices and therefore of profits of the said machinery and raw materials, on sales which have already occurred. What critics present as an inconsistency between 'values' and 'prices of production' is simply a recognition of *two different time-frameworks* (cycles) in which the equalization of the rate of profit has been achieved, a first one for inputs, and a second, later one for outputs.

Marx's Theory of Rent

The labour theory of value defines value as the socially necessary quantity of labour determined by the average productivity of labour of each given sector of production. But these values are not mathematically fixed data. They are simply the expression of a *process* going on in real life, under capitalist commodity production. So this average is only ascertained in the course of a certain time-span. There is a lot of logical argument and empirical evidence to advance the hypothesis that the normal time-span for essentially modifying the value of commodities is the business cycle, from one crisis of over-production (recession) to the next one.

Before technological progress and (or) better (more 'rational') labour organization etc. determines a more than marginal change (in general: decline) in the value of a commodity, and the crisis eliminates less efficient firms, there will be a coexistence of firms with various 'individual values' of a given commodity in a given branch of output, even assuming a single market price. So, in his step-for-step approach towards explaining the immediate phenomena (facts of economic life) like prices and profits, by their essence, Marx introduces at this point of his analysis a new mediating concept, that of *market value* (Marx, *g*, ch. 10). The market value of a commodity is the 'individual value' of the firm, or a group of firms, in a given branch of production, around which the market price will fluctuate. That 'market value' is not necessarily the mathematical (weighted) average of labour expenditure of all firms of that branch. It can be below, equal or above that average, for a certain period (generally less than the duration of the business cycle, at least under 'free competition'), according to whether social demand is saturated, just covered or to an important extent not covered by current output plus existing stocks. In these three cases respectively, the more (most) efficient firms, the firms of average efficiency, or even firms with labour productivity below average, will determine the market value of that given commodity.

This implies that the more efficient firms enjoy *surplus profits* (profits over and above the average profit) in case 2 and 3 and that a certain number of firms work at less than average profit in all three cases, but especially in case 1.

The mobility of capital, i.e. normal capitalist competition, generally eliminates such situations after a certain lapse of time. But when that mobility of capital is impeded for long periods by either unavoidable scarcity (natural conditions not renewable or non-substitutable, like land and mineral deposits) or through the operation of institutional obstacles (private property of land and mineral resources forbidding access to available capital, except in exchange for payments over and above average profit), these surplus profits can be frozen and maintained for decades. They thus become *rents,* of which *ground rent* and *mineral rent* are the most obvious examples in Marx's time, extensively analysed in *Capital* vol. III (Marx, *g*, part 6).

Marx's theory of rent is the most difficult part of his economic theory, the one which has witnessed fewer comments and developments, by followers and critics alike, than other major parts of his 'system'. But it is not obscure. And in contrast to Ricardo's or Rodbertus's theories of rent, it represents a straightforward application of the labour theory of value. It does not imply any emergence of ' supplementary' value (surplus value, profits) in the market, in the process of

circulation of commodities, which is anathema to Marx and to all consistent upholders of the labour theory of value. Nor does it in any way suggest that land or mineral deposits 'create' value.

It simply means that in agriculture and mining less productive labour (as in the general case analysed above) determines the market value of food or minerals, and that therefore more efficient farms and mines enjoy surplus profits which Marx calls differential (land and mining) rent. It also means that as long as productivity of labour in agriculture is generally below the average of the economy as a whole (or more correctly: that the organic composition of capital, the expenditure in machinery and raw materials as against wages, is inferior in agriculture to that of industry and transportation), the sum-total of surplus-value produced in agriculture will accrue to landowners + capitalist farmers taken together, and will not enter the general process of (re)distribution of profit throughout the economy as a whole.

This creates the basis for a supplementary form of rent, over and above differential rent, rent which Marx calls absolute land rent. This is, incidentally, the basis for a long-term separation of capitalist landowners from entrepreneurs in farming or animal husbandry, distinct from feudal or semi-feudal landowners or great landowners under conditions of predominantly petty commodity production, or in the Asiatic mode of production, with free peasants.

The validity of Marx's theory of land and mining rents has been confirmed by historical evidence, especially in the 20th century. Not only has history substantiated Marx's prediction that, in spite of the obstacle of land and mining rent, mechanization would end up by penetrating food and raw materials production too, as it has for a long time dominated industry and transportation, thereby causing a growing decline of differential rent (this has occurred increasingly in agriculture in the last 25–50 years, first in North America, and then in Western Europe and even elsewhere). It has also demonstrated that once the structural scarcity of food disappears, the institutional obstacle (private property) loses most of its efficiency as a brake upon the mobility of capital. Therefore the participation of surplus-value produced in agriculture in the general process of profit equalization throughout the economy cannot be prevented any more. Thereby absolute rent tends to wither away and, with it, the separation of land ownership from entrepreneurial farming and animal husbandry. It is true that farmers can then fall under the sway of the banks, but they do so as private owners of their land which becomes mortgaged, not as share-croppers or entrepreneurs renting land from separate owners.

On the other hand, the reappearance of structural scarcity in the realm of energy enabled the OPEC countries to multiply the price of oil by ten in the 1970s, i.e. to have it determined by the oilfields where production costs are the highest, thereby assuring the owners of the cheapest oil wells in Arabia, Iran, Libya, etc. of huge differential mineral rents.

Marx's theory of land and mineral rent can be easily extended into a general theory of rent, applicable to all fields of production where formidable difficulties of entry limit mobility of capital for extended periods of time. It thereby becomes the basis of a Marxist theory of monopoly and monopoly surplus profits, i.e. in the form of cartel rents (Hilferding 1910) or of technological rent (Mandel 1972). Lenin's and Bukharin's theories of surplus profit are based upon analogous but not identical reasoning (Bukharin 1914, 1926; Lenin 1917).

But in all these cases of general application of the Marxist theory of rent, the same caution should apply as Marx applied to his theory of land rent. By its very nature, capitalism, based upon private property, i.e. 'many capitals' – that is, competition – cannot tolerate any 'eternal' monopoly, a 'permanent' surplus profit deducted from the sum total of profits which is divided among the capitalist class as a whole. Technological innovations, substitution of new products for old ones including the field of raw materials and of food, will in the long run reduce or eliminate all monopoly situations, especially if the profit differential is large enough to justify huge research and investment outlays.

Marx's Theory of Money

In the same way as his theory of rent, Marx's theory of money is a straightforward application of the labour theory of value. As value is but the embodiment of socially necessary labour, commodities exchange with each other in proportion of the labour quanta they contain. This is true for the exchange of iron against wheat as it is true for the exchange of iron against gold or silver. Marx's theory of money is therefore in the first place a commodity theory of money. A given commodity can play the role of universal medium of exchange, as well as fulfil all the other functions of money, precisely because it is a commodity, i.e. because it is itself the product of socially necessary labour. This applies to the precious metals in the same way it applies to all the various commodities which, throughout history, have played the role of money.

It follows that strong upheavals in the 'intrinsic' value of the money-commodity will cause strong upheavals in the general price level. In Marx's theory of money, (market) prices are nothing but the expression of the value of commodities in the value of the money commodity chosen as a monetary standard. If £1 sterling $= \frac{1}{10}$ ounce of gold, the formula 'the price of 10 quarters of wheat is £1' means that 10 quarters of wheat have been produced in the same socially necessary labour time as $\frac{1}{10}$ ounce of gold. A strong decrease in the average productivity of labour in gold mining (as a result for example of a depletion of the richer gold veins) will lead to a general depression of the average price level, all other things remaining equal. Likewise, a sudden and radical increase in the average productivity of labour in gold mining, through the discovery of new rich gold fields (California after 1848; the Rand in South Africa in the 1890s) or through the application of new revolutionary technology, will lead to a general increase in the price level of all other commodities.

Leaving aside short-term oscillations, the general price level will move in medium and long-term periods according to the relation between the fluctuations of the productivity of labour in agriculture and industry on the one hand, and the fluctuations of the productivity of labour in gold mining (if gold is the money- commodity), on the other.

Basing himself on that commodity theory of money. Marx therefore criticized as inconsistent Ricardo's quantity theory (Marx, *h*, part 2). But for exactly the same reason of a consistent application of the labour theory of value, the quantity of money in circulation enters Marx's economic analysis when he deals with the phenomenon of paper money (Marx, *c*).

As gold has an intrinsic value, like all other commodities, there can be no 'gold inflation', as little as there can be a 'steel inflation'. Abstraction made of short-term price fluctuations caused by fluctuations between supply and demand, a persistent decline of the value of gold (exactly as for all other commodities) can only be the result of a persistent increase in the average productivity of labour in gold mining, and not of an 'excess' of circulation in gold. If the demand for gold falls consistently, this can only indirectly trigger off a decline in the value of gold through causing the closure of the least productive gold mines. But in the case of the money-commodity, such overproduction can hardly occur, given the special function of gold of serving as a universal reserve fund, nationally and internationally. It will always therefore find a buyer, be it not, of course, always at the same 'prices' (in Marx's economic theory, the concept of 'price of gold' is meaningless. As the price of a commodity is precisely its expression in the value of gold, the 'price of gold' would be the expression of the value of gold in the value of gold).

Paper money, bank notes, are a *money sign* representing a given quantity of the money-commodity. Starting from the above-mentioned example, a banknote of £1 represents $\frac{1}{10}$ ounce of gold. This is an objective 'fact of life', which no government or monetary authority can arbitrarily alter. It follows that any emission of paper money in excess of that given proportion will automatically lead to an increase in the general price level, always other things remaining equal. If £1 suddenly represents only $\frac{1}{20}$ ounce of gold, because

paper money circulation has doubled without a significant increase in the total labour time spent in the economy, then the price level will tend to double too. The value of $\frac{1}{10}$, ounce of gold remains equal to the value of 10 quarters of wheat. But as $\frac{1}{10}$, ounce of gold is now represented by £2 in paper banknotes instead of being represented by £1, the price of wheat will move from £1 to £2 for 10 quarters (from two shillings to four shillings a quarter before the introduction of the decimal system).

This does not mean that in the case of paper money, Marx himself has become an advocate of a quantity theory of money. While there are obvious analogies between his theory of paper money and the quantity theory, the main difference is the rejection by Marx of any mechanical automatism between the quantity of paper money emitted on the one hand, and the general dynamic of the economy (including on the price level) on the other.

In Marx's explanation of the movement of the capitalist economy in its totality, the formula *ceteris paribus* is meaningless. Excessive (or insufficient) emission of paper money never occurs in a vacuum. It always occurs at a given stage of the business cycle, and in a given phase of the longer-term historical evolution of capitalism. It is thereby always combined with given ups and downs of the rate of profit, of productivity of labour, of output, of market conditions (overproduction or insufficient production). Only in connection with these other fluctuations can the effect of paper money 'inflation' or 'deflation' be judged, including the effect on the general price level. The key variables are in the field of production. The key synthetic resultant is in the field of profit. Price movements are generally epiphenomena as much as they are signals. To untwine the tangle, more is necessary than a simple analysis of the fluctuations of the quantity of money. Only in the case of extreme runaway inflation of paper money would this be otherwise; and even in that border case, *relative* price movements (different degrees of price increases for different commodities) would still confirm that, in the last analysis, the law of value rules, and not the arbitrary decisions of the Central Bank, or any other authority controlling or emitting paper money.

Marx's Theory of Surplus-Value

Marx himself considered his theory of surplus-value his most important contribution to the progress of economic analysis (Marx, *I*; letter to Engels of 24 August 1867). It is through this theory that the wide scope of his sociological and historical thought enables him simultaneously to place the capitalist mode of production in its historical context, and to find the roots of its inner economic contradictions and its laws of motion in the specific relations of production on which it is based.

As said before, Marx's theory of classes is based on the recognition that in each class society, part of society (the ruling class) appropriates the social surplus product. But that surplus product can take three essentially different forms (or a combination of them). It can take the form of straightforward unpaid surplus labour, as in the slave mode of production, early feudalism or some sectors of the Asian mode of production (unpaid corvée labour for the Empire). It can take the form of goods appropriated by the ruling class in the form of use-values pure and simple (the products of surplus labour), as under feudalism when feudal rent is paid in a certain amount of produce (produce rent) or in its more modern remnants, such as sharecropping. And it can take a money form, like money-rent in the final phases of feudalism, and capitalist profits. Surplus-value is essentially just that: the money form of the social surplus produce or, what amounts to the same, the money product of surplus labour. It has therefore a common root with all other forms of surplus product: unpaid labour.

This means that Marx's theory of surplus-value is basically a deduction (or residual) theory of the ruling classes' income. The whole social product (the net national income) is produced in the course of the process of production, exactly as the whole crop is harvested by the peasants. What happens on the market (or through appropriation of the produce) is a distribution (or redistribution) of what already has been created. The surplus product, and therefore also its money form, surplus-value, is the residual of that new (net) social product (income) which remains after the producing classes have received their compensation

(under capitalism: their wages). This 'deduction' theory of the ruling classes' income is thus *ipso facto* an exploitation theory. Not in the ethical sense of the word – although Marx and Engels obviously manifested a lot of understandable moral indignation at the fate of all the exploited throughout history, and especially at the fate of the modern proletariat – but in the economical one. The income of the ruling classes can always be reduced in the final analysis to the product of unpaid labour: that is the heart of Marx's theory of exploitation.

That is also the reason why Marx attached so much importance to treating *surplus-value as a general category*, over and above profits (themselves subdivided into industrial profits, bank profits, commercial profits etc.), interest and rent, which are all part of the total surplus product produced by wage labour. It is this general category which explains both the existence (the common interest) of the ruling class (all those who live off surplus value), and the origins of the class struggle under capitalism.

Marx likewise laid bare the economic mechanism through which surplus-value originates. As the basis of that economic mechanism is a huge social upheaval which started in Western Europe in the 15th century and slowly spread over the rest of the continent and all other continents (in many so-called underdeveloped countries, it is still going on to this day).

Through many concomitant economic (including technical), social, political and cultural transformations, the mass of the direct producers, essentially peasants and handicraftsmen, are separated from their means of production and cut off from free access to the land. They are therefore unable to produce their livelihood on their own account. In order to keep themselves and their families alive, they have to hire out their arms, their muscles and their brains, to the owners of the means of production (including land). If and when these owners have enough money capital at their disposal to buy raw materials and pay wages, they can start to organize production on a capitalist basis, using wage labour to transform the raw materials which they buy, with the tools they own, into finished products which they then automatically own too.

The capitalist mode of production thus presupposes that the producers' *labour power has become a commodity.* Like all other commodities, the commodity labour power has an exchange value and a use value. The exchange value of labour power, like the exchange value of all other commodities, is the amount of socially necessary labour embodied in it, i.e. its reproduction costs. This means concretely the value of all the consumer goods and services necessary for a labourer to work day after day, week after week, month after month, at approximately the same level of intensity, and for the members of the labouring classes to remain approximately stable in number and skill (i.e. for a certain number of working-class children to be fed, kept and schooled, so as to replace their parents when they are unable to work any more, or die). But the use value of the commodity labour power is precisely its capacity to create new value, including its potential to create more value than its own reproduction costs. Surplus-value is but that difference between the total new value created by the commodity labour power, and its own value, its own reproduction costs.

The whole Marxian theory of surplus-value is therefore based upon that subtle distinction between 'labour power' and 'labour' (or value). But there is nothing 'metaphysical' about this distinction. It is simply an explanation (demystification) of a process which occurs daily in millions of cases.

The capitalist does not buy the worker's 'labour'. If he did that there would be obvious theft, for the worker's wage is obviously smaller than the total value he adds to that of the raw materials in the course of the process of production. No: the capitalist buys 'labour power', and often (not always of course) he buys it at its *justum pretium*, at its real value. So he feels unjustly accused when he is said to have caused a 'dishonest' operation. The worker is victim not of vulgar theft but of a social set-up which condemns him first to transform his productive capacity into a commodity, then to sell that labour power on a specific market (the labour market) characterized by institutional inequality, and finally to content himself with the market price he can get for that commodity, irrespective of whether the new value

he creates during the process of production exceeds that market price (his wage) by a small amount, a large amount, or an enormous amount.

The labour power the capitalist has bought 'adds value' to that of the used-up raw materials and tools (machinery, buildings etc.). If, and until that point of time, this added value is inferior or equal to the workers' wages, surplus-value cannot originate. But in that case, the capitalist has obviously no interest in hiring wage labour. He only hires it because that wage labour has the quality (the use value) to add to the raw materials' value more than its own value (i.e. its own wages). This 'additional added value' (the difference between total 'value added' and wages) is precisely surplus-value. Its emergence from the process of production is the precondition for the capitalists' hiring workers, for the existence of the capitalist mode of production.

The institutional inequality existing on the labour market (masked for liberal economists, sociologists and moral philosophers alike by juridical equality) arises from the very fact that the capitalist mode of production is based upon generalized commodity production, generalized market economy. This implies that a propertyless labourer, who owns no capital, who has no reserves of larger sums of money but who has to buy his food and clothes, pay his rent and even elementary public transportation for journeying between home and workplace, *in a continuous way* in exchange of money, is under the *economic compulsion* to sell the only commodity he possesses, to wit his labour power, also on a continuous basis. He cannot withdraw from the labour market until the wages go up. He cannot wait.

But the capitalist, who has money reserves, can temporarily withdraw from the labour market. He can lay his workers off, can even close or sell his enterprise and wait a couple of years before starting again in business. This institutional difference makes price determination of the labour market a game with loaded dice, heavily biased against the working class. One just has to imagine a social set-up in which each citizen would be guaranteed an annual minimum income by the community, irrespective or whether he is employed or not, to understand that 'wage determination' under these circumstances would be quite different from what it is under capitalism. In such a set-up the individual would really have the economic choice whether to sell his labour power to another person (or a firm) or not. Under capitalism, he has no choice. His is forced by economic compulsion to go through with that sale, practically at any price.

The economic function and importance of trade unions for the wage-earners also clearly arises from that elementary analysis. For it is precisely the workers' 'combination' and their assembling a collective resistance fund (what was called by the first French unions *caisses de résistance*, 'reserve deposits') which enables them, for example though a strike, to withdraw the supply of labour power temporarily from the market so as to stop a downward trend of wages or induce a wage increase. There is nothing 'unjust' in such a temporary withdrawal of the supply of labour power, as there are constant withdrawals of demand for labour power by the capitalists, sometimes on a huge scale never equalled by strikes. Through the functioning of strong labour unions, the working class tries to correct, albeit partially and modestly, the institutional inequality on the labour market of which it is a victim, without ever being able to neutralize it durably or completely.

It cannot neutralize it durably because in the very way in which capitalism functions there is a powerful built-in corrective in favour of capital: the inevitable emergence of an industrial reserve army of labour. There are three key sources for that reserve army: the mass of precapitalist producers and self-employed (independent peasants, handicraftsmen, trades-people, professional people, small and medium-sized capitalists); the mass of housewives (and to a lesser extent, children); the mass of the wage-earners themselves, who potentially can be thrown out of employment.

The first two sources have to be visualized not only in each capitalist country seen separately but on a world scale, through the operations of international migration. They are still unlimited to a great extent, although the number of wage-earners the world over (including agricultural wage labourers) has already passed the one billion mark. At the third source, while it is obviously

not unlimited (if wage labour would disappear altogether, if all wage labourers would be fired, surplus- value production would disappear too; that is why 'total robotism' is impossible under capitalism), its reserves are enormous, precisely in tandem with the enormous growth of the absolute number of wage earners.

The fluctuations of the industrial reserve army are determined both by the business cycle and by long-term trends of capital accumulation. Rapidly increasing capital accumulation attracts wage labour on a massive scale, including through international migration. Likewise, deceleration, stagnation or even decline of capital accumulation inflates the reserve army of labour. There is thus an upper limit to wage increases, when profits (realized profits and expected profits) are 'excessively' reduced in the eyes of the capitalists, which triggers off such decelerated, stagnating or declining capital accumulation, thereby decreasing employment and wages, till a 'reasonable' level of profits is restored.

This process does not correspond to any 'natural economic law' (or necessity), nor does it correspond to any 'immanent justice'. It just expresses the inner logic of the *capitalist* mode of production, which is geared to profit. Other forms of economic organization could function, have functioned and are functioning on the basis of other logics, which do not lead to periodic massive unemployment. On the contrary, a socialist would say — and Marx certainly thought so — that the capitalist system is an 'unjust', or better stated 'alienating', 'inhuman' social system, precisely because it cannot function without periodically reducing employment and the satisfaction of elementary needs for tens of millions of human beings.

Marx's theory of surplus-value is therefore closely intertwined with a *theory of wages* which is far away from Malthus's, Ricardo's or the early socialists' (like Ferdinand Lassalle's) 'iron law of wages', in which wages tend to fluctuate around the physiological minimum. That crude theory of 'absolute pauperization' of the working class under capitalism, attributed to Marx by many authors (Popper 1945, et al.), is not Marx's at all, as many contemporary authors have convincingly demonstrated (see among others Rosdolsky 1968). Such an 'iron law of wages' is essentially a demographic one, in which birth rates and the frequency of marriages determine the fluctuation of employment and unemployment and thereby the level of wages.

The logical and empirical inconsistencies of such a theory are obvious. Let it be sufficient to point out that while fluctuations in the supply of wage-labourers are considered essential, fluctuations in the demand for labour power are left out of the analysis. It is certainly a paradox that the staunch opponent of capitalism, Karl Marx, pointed out already in the middle of the 19th century the potential for wage increases under capitalism, even though not unlimited in time and space. Marx also stressed the fact that for each capitalist wage increases of other capitalists' workers are considered increases of potential purchasing power, not increases in costs (Marx, *d*).

Marx distinguishes two parts in the workers' wage, two elements of reproduction costs of the commodity labour power. One is purely physiological, and can be expressed in calories and energy quanta; this is the bottom below which the wage cannot fall without destroying slowly or rapidly the workers' labour capacity. The second one is historical-moral, as Marx calls it (Marx, *i*), and consists of those additional goods and services which a shift in the class relationship of forces, such as a victorious class struggle, enables the working class to incorporate into the average wage, the socially necessary (recognized) reproduction costs of the commodity labour power (e.g. paid holidays after the French general strike of June 1936). This part of the wage is essentially flexible. It will differ from country to country, continent to continent, and from epoch to epoch, according to many variables. But it has the upper limit indicated above: the ceiling from which profits threaten to disappear, or to become insufficient in the eyes of the capitalists, who then go on an 'investment strike'.

So Marx's theory of wages is essentially an *accumulation-of-capital theory of wages* which sends us back to what Marx considered the first 'law of motion' of the capitalist mode of production: the compulsion for the capitalists to step up constantly the rate of capital accumulation.

The Laws of Motion of the Capitalist Mode of Production

Marx's theory of surplus-value is his most revolutionary contribution to economic science, his discovery of the basic long-term 'laws of motion' (development trends) of the capitalist mode of production constitutes undoubtedly his most impressive scientific achievement. No other 19th-century author has been able to foresee in such a coherent way how capitalism would function, would develop and would transform the world, as did Karl Marx. Many of the most distinguished contemporary economists, starting with Wassily Leontief (1938), and Joseph Schumpeter, (1942) have recognized this.

While some of these 'laws of motion' have obviously created much controversy, we shall nevertheless list them in logical order, rather than according to the degree of consensus they command.

(a) *The capitalist's compulsion to accumulate.* Capital appears in the form of accumulated money, thrown into circulation in order to increase in value. No owner of money capital will engage in business in order to recoup exactly the sum initially invested, and nothing more than that. By definition, the search for profit is at the basis of all economic operations by owners of capital.

Profit (surplus-value, accretion of value) can originate outside the sphere of production in a precapitalist society. It represents then essentially a *transfer of value* (so-called primitive accumulation of capital); but under the capitalist mode of production, in which capital has penetrated the sphere of production and dominates it, surplus-value is currently produced by wage labour. It represents a constant increase in value.

Capital can only appear in the form of many capitals, given its very historical- social origin in private property (appropriation) of the means of production. 'Many capitals' imply unavoidable competition. Competition in a capitalist mode of production is competition for selling commodities in an anonymous market. While surplus-value is produced in the process of production, it is *realized* in the process of circulation, i.e. through the sale of the commodities. The capitalist wants to sell at maximum profit. In practice, he will be satisfied if he gets the average profit, which is a percentage really existing in his consciousness (e.g. Mr. Charles Wilson, the then head of the US automobile firm General Motors, stated before a Congressional enquiry: we used to fix the expected sales price of our cars by adding 15% to production costs). But he can never be sure of this. He cannot even be sure that all the commodities produced will find a buyer.

Given these uncertainties, he has to strive constantly to get the better of his competitors. This can only occur through operating with more capital. This means that at least part of the surplus-value produced will not be unproductively consumed by the capitalists and their hangers-on through luxury consumption, but will be accumulated, added to the previously existing capital.

The inner logic of capitalism is therefore not only to 'work for profit', but also to 'work for capital accumulation'. 'Accumulate, accumulate; that is Moses and the Prophets', states Marx in *Capital,* Vol. I (Marx, *e*, p. 742). Capitalists are *compelled* to act in that way as a result of competition. It is competition which basically fuels this terrifying snowball logic: initial value of capital → accretion of value (surplus- value) → accretion of capital → more accretion of surplus-value → more accretion of capital etc. 'Without competition, the fire of growth would burn out' (Marx, *g*, p. 368).

(b) *The tendency towards constant technological revolutions.* In the capitalist mode of production, accumulation of capital is in the first place accumulation of productive capital, or capital invested to produce more and more commodities. Competition is therefore above all competition between productive capitals, i.e. 'many capitals' engaged in mining, manufacturing, transportation, agriculture, telecommunications. The main weapon in competition between capitalist firms is cutting

production costs. More advanced production techniques and more 'rational' labour organization are the main means to achieve that purpose. The basic tend of capital accumulation in the capitalist mode of production is therefore a trend towards more and more sophisticated machinery. Capitalist growth takes the dual form of higher and higher value of capital and of constant revolutions in the techniques of production, of constant technological progress.

(c) *The capitalists' unquenchable thirst for surplus-value extraction.* The compulsion for capital to grow, the irresistible urge for capital accumulation, realizes itself above all through a constant drive for the increase of the production of surplus- value. Capital accumulation is nothing but surplus-value capitalization, the transformation of part of the new surplus-value into additional capital. There is no other source of additional capital than additional surplus-value produced in the process of production.

Marx distinguishes two different forms of additional surplus-value production. *Absolute surplus-value* accretion occurs essentially through the extension of the work day. If the worker reproduces the equivalent of his wages in 4 hours a day, an extension of the work day from 10 to 12 hours will increase surplus-value from 6 to 8 hours. *Relative surplus-value* accretion occurs through an increase of the productivity of labour in the wage-goods sector of the economy. Such an increase in productivity implies that the equivalent of the value of an identical basket of goods and services consumed by the worker could be produced in 2 hours instead of 4 hours of labour. If the work day remains stable at 10 hours and real wages remain stable too, surplus-value will then increase from 6 to 8 hours.

While both processes occur throughout the history of the capitalist mode of production (viz. the contemporary pressure of employers in favour of overtime!), the first one was prevalent first, the second one became prevalent since the second half of the 19th century, first in Britain, France and Belgium, then in the USA and Germany, later in the other industrialized capitalist countries, and later still in the semi-industrialized ones. Marx calls this process the *real subsumption* (subordination) of *labour under capital* (Marx, *k*), for it represents not only an economic but also a physical subordination of the wage-earner under the machine. This physical subordination can only be realized through social control. The history of the capitalist mode of production is therefore also the history of successive forms of – tighter and tighter – control of capital over the workers inside the factories (Braverman 1974); and of attempts at realizing that tightening of control in society as a whole.

The increase in the production of relative surplus-value is the goal for which capitalism tends to periodically substitute machinery for labour, i.e. to expand the industrial reserve army of labour. Likewise, it is the main tool for maintaining a modicum of social equilibrium, for when productivity of labour strongly increases, above all in the wage-good producing sectors of the economy, real wages and profits (surplus-value) can both expand simultaneously. What were previously luxury goods can even become mass-produced wage goods.

(d) *The tendency towards growing concentration and centralization of capital.* The growth of the value of capital means that each successful capitalist firm will be operating with more and more capital. Marx calls this the tendency towards growing concentration of capital. But in the competitive process, there are victors and vanquished. The victors grow. The vanquished go bankrupt or are absorbed by the victors. This process Marx calls the centralization of capital. It results in a declining number of firms which survive in each of the key fields of production. Many small and medium-sized capitalists disappear as independent business men and women. They become in turn salary earners, employed by successful capitalist firms. Capitalism itself is the big 'expropriating' force, suppressing

private property of the means of production for many, in favour of private property for few.

(e) *The tendency for the 'organic composition of capital' to increase.* Productive capital has a double form. It appears in the form of *constant* capital: buildings, machinery, raw materials, energy. It appears in the form of *variable* capital: capital spent on wages of productive workers. Marx calls the part of capital used in buying labour power variable, because only that part produces additional value. In the process of production, the value of constant capital is simply maintained (transferred *in toto* or in part into the value of the finished product). Variable capital on the contrary is the unique source of 'added value'.

Marx postulates that the basic historic trend of capital accumulation is to increase investment in constant capital at a quicker pace than investment in variable capital; the relation between the two he calls the 'organic composition of capital'. This is both a technical/physical relation (a given production technique implies the use of a given number of productive wage earners, even if not in an absolutely mechanical way) and a value relation. The trend towards an increase in the 'organic composition of capital' is therefore a historical trend towards *basically labour-saving technological progress.*

This tendency has often been challenged by critics of Marx. Living in the age of semi-automation and 'robotism', it is hard to understand that challenge. The conceptual confusion on which this challenge is mostly based is an operation with the 'national wage bill', i.e. a confusion between wages in general and variable capital, which is only the wage bill of productive labour. A more correct index would be the part of the labour costs in total production costs in the manufacturing (and mining) sector. It is hard to deny that this proportion shows a downward secular trend.

(f) *The tendency of the rate of profit to decline.* For the workers, the basic relation they are concerned with is the rate of surplus-value, i.e. the division of 'value added' by them between wages and surplus-value. When this goes up, their exploitation (the unpaid labour they produce) obviously goes up. For the capitalists however, this relationship is not meaningful. They are concerned with the relation between surplus-value and the *totality* of capital invested, never mind whether in the form of machinery and raw materials or in the form of wages. This relation is the *rate of profit*. It is a function of two variables, the organic composition of capital and the rate of surplus-value. If the value of constant capital is represented by c, the value of variable capital (wages of productive workers) by v and surplus-value by s, the rate of profit will be $s/(c + v)$. This can be rewritten as

$$\frac{s/v}{(c+v)/(v)} + 1$$

with the two variables emerging ($(c + v)/(v)$ obviously reflects c).

Marx postulates that the increase in the rate of surplus value has definite limits, while the increase in the organic composition of capital has practically none (automation, robotism). There will therefore be a basic tendency for the rate of profit to decline.

This is however absolutely true only on a very long-term, i.e. essentially 'secular', basis. In other time-frameworks, the rate of profit can fluctuate under the influence of countervailing forces. Constant capital can be devalorized, through 'capital saving' technical process, and through economic crises (see below). The rate of surplus-value can be strongly increased in the short or medium term, although each strongly increase makes a further increase more difficult (Marx, *d*, pp. 335–6); and capital can flow to countries (e.g. 'Third World' ones) or branches (e.g. service sectors) where the organic composition of capital is significantly lower than in the previously industrialized ones, thereby raising the average rate of profit.

Finally, the increase in the mass of *surplus-value* – especially through the extension of wage labour in general, i.e. the total number of workers – offsets to a large extent the depressing

effects of moderate declines of the average rate of profit. Capitalism will not go out of business if the mass of surplus-value produced increases 'only' from £10 to £17 billion, while the total mass of capital has moved from 100 to 200 billion; and capital accumulation will not stop under these circumstances, nor necessarily slow down significantly. It would be sufficient to have the unproductively consumed part of surplus-value pass e.g. from £3 to £2 billion, to obtain a rate of capital accumulation of 15/200, i.e. 7.5%, even higher than the previous one of 7/100, in spite of a decline of the rate of profit from 10 to 8.5%.

(g) *The inevitability of class struggle under capitalism.* One of the most impressive projections by Marx was that of the inevitability of elementary class struggle under capitalism. Irrespective of the social global framework or of their own historical background, wage-earners will fight everywhere for higher real wages and a shorter work day. They will form elementary organizations for the collective instead of the individual sale of the commodity labour power, i.e. trade unions.

While at the moment Marx made that projection there were less than half a million organized workers in at the most half a dozen countries in the world, today trade unions encompass hundreds of millions of wage-earners spread around the globe. There is no country, however remote it might be, where the introduction of wage labour has not led to the appearance of workers' coalitions.

While elementary class struggle and elementary unionization of the working class are inevitable under capitalism, higher, especially political forms of class struggle, depend on a multitude of variables as to the rapidity with which they extend beyond smaller minorities of each 'national' working class and internationally. But there too the basic secular trend is clear. There were in 1900 innumerably more conscious socialists than in 1850, fighting not only for better wages but, to use Marx's words, for the abolition of wage labour (Marx, *i*) and organizing working class parties for that purpose. There are today many more than in 1900.

(h) *The tendency towards growing social polarization.* From two previously enumerated trends, the trend towards growing centralization of capital and the trend towards the growth of the mass of surplus-value, flows the trend towards growing social polarization under capitalism. The proportion of the active population represented by wage-labour in general, i.e. by the modern proletariat (which extends far beyond productive workers in and by themselves) increases. The proportion represented by self-employed (small, medium-sized and big capitalists, as well as independent peasants, handicraftsmen, tradespeople and 'free professions' working without wage-labour) decreases. In fact, in several capitalist countries, the first category has already passed the 90 per cent mark, while in Marx's time it was below 50 per cent everywhere but in Britain. In most industrialized (imperialist) countries, it has reached 80 – 85 per cent.

This does not mean that the petty entrepreneurs have tended to disappear. Ten or 15 – 20 per cent out of 30 million people, not to say out of 120 million, still represent a significant social layer. While many small businesses disappear, especially in times of economic depression, as a result of severe competition, they also are constantly created, especially in the interstices between big firms, and in new sectors where they play an exploratory role. Also, the overall social results of growing proletarization are not simultaneous with the economic process in and by itself. From the point of view of class consciousness, culture, political attitude, there can exist significant time-lags between the transformation of an independent farmer, grocer or doctor into a wage-earner, and his acceptance of socialism as an overall social solution for his own and society's ills. But again, the secular trend is towards *growing homogeneity,* less and less heterogeneity, of the mass of the wage-earning class, and not the other way around. It is sufficient to compare the differences in consumer patterns, attitudes towards unionization or voting habits between manual workers, bank employees and government functionaries in say 1900 and

today, to note that they have decreased and not increased.

(i) *The tendency towards growing objective socialization of labour.* Capitalism starts in the form of private production on a medium-sized scale for a limited number of largely unknown customers, on an uncontrollably wide market, i.e. under conditions of near complete fragmentation of social labour and anarchy of the economic process. But as a result of growing technological progress, tremendously increased concentration of capital, the conquest of wider and wider markets throughout the world, and the very nature of the labour organization inside large and even medium-sized capitalist factories, a powerful process of objective socialization of labour is simultaneously set in motion. This process constantly extends the sphere of economy in which not blind market laws by conscious decisions and even large-scale cooperation prevail.

This is true especially inside mammoth firms (inside multinational corporations, such 'planning' prevails far beyond the boundaries of nation-states, even the most powerful ones!) and inside large-scale factories; but it is also increasingly true for buyer/seller relations, in the first place on an inter-firm basis, between public authorities and firms, and more often than one thinks between traders and consumers too. In all these instances, the rule of the law of value becomes more and more remote, indirect and discontinuous. Planning prevails on a short and even medium- term basis.

Certainly, the economy still remains capitalist. The rule of the law of value imposes itself brutally through the outburst of economic crises. Wars and social crises are increasingly added to these economic crises to remind society that, under capitalism, this growing objective socialization of labour and production is indissolubly linked to private appropriation, i.e. to the profit motive as motor of economic growth. That linkage makes the system more and more crisis-ridden; but at the same time the growing socialization of labour and production creates the objective basis for a general socialization of the economy, i.e. represents the basis of the coming socialist order created by capitalism itself, within the framework of its own system.

(j) *The inevitability of economic crises under capitalism.* This is another of Marx's projections which has been strikingly confirmed by history. Marx ascertained that periodic crises of overproduction were unavoidable under capitalism. In fact, since the crisis of 1825, the first one occurring on the world market for industrial goods to use Marx's own formula, there have been twenty-one business cycles ending (or beginning, according to the method of analysis and measurement used) with twenty-one crises of overproduction. A twenty-second is appearing on the horizon as we are writing.

Capitalist economic crises are always crises of *overproduction of commodities (exchange values)*, as opposed to pre- and post-capitalist economic crises, which are essentially crises of *underproduction of use-values.* Under capitalist crises, *expanded reproduction* – economic growth – is brutally interrupted, not because too few commodities have been produced but, on the contrary, because a mountain of produced commodities finds no buyers. This unleashes a spiral movement of collapse of firms, firing of workers, contraction of sales (or orders) for raw materials and machinery, new redundancies, new contraction of sales of consumer goods etc. Through this *contracted reproduction*, prices (gold prices) collapse, production and income is reduced, capital loses value. At the end of the declining spiral, output (and stocks) have been reduced more than purchasing power. Then production can pick up again; and as the crisis has both increased the rate of surplus-value (through a decline of wages and a more 'rational' labour organization) and decreased the value of capital, the average rate of profit increases. This stimulates investment. Employment increases, value production and national income expand, and we enter a new cycle of economic revival, prosperity, overheating and the next crisis.

No amount of capitalists' (essentially large combines' and monopolies') 'self- regulation', no amount of government intervention, has been able to suppress this cyclical movement of capitalist production. Nor can they succeed in achieving that result. This cyclical movement is inextricably linked to production for profit and private property (competition), which imply periodic over-shooting (too little or too much investment and output), precisely because each firm's attempt at maximizing profit unavoidably leads to a lower rate of profit for the system as a whole. It is likewise linked to the separation of value production and value realization.

The only way to avoid crises of overproduction is to eliminate all basic sources of disequilibrium in the economy, including the disequilibrium between productive capacity and purchasing power of the 'final consumers'. This calls for elimination of generalized commodity production, of private property and of class exploitation, i.e. for the elimination of capitalism.

Marx's Theory of Crises

Marx did not write a systematic treatise on capitalist crises. His major comments on the subject are spread around his major economic writings, as well as his articles for the *New York Daily Tribune.* The longest treatment of the subject is in his *Theorien über den Mehrwert,* subpart on Ricardo (Marx, *h*, Part 2). Starting from these profound but unsystematic remarks, many interpretation of the 'Marxist theory or crisis' have been offered by economists who consider themselves Marxists. 'Monocausal' ones generally centre around 'disproportionality' (Bukharin, Hilferding, Otto Bauer) – anarchy of production as the key cause of crises – or 'underconsumption' – lack of purchasing power of the 'final consumers' as the cause of crises (Rosa Luxemburg, Sweezy). 'Non-monocausal' ones try to elaborate Marx's own *dictum* according to which *all* basic contradictions of the capitalist mode of production come into play in the process leading to a capitalist crisis (Grossman, Mandel).

The question of determining whether according to Marx, a crisis of overproduction is first of all a crisis of overproduction of commodities or a crisis of overproduction of capital is really meaningless in the framework of Marx's economic analysis. The mass of commodities is but one specific form of capital, commodity capital. Under capitalism, which is generalized commodity production, no overproduction is possible which is not simultaneously overproduction of commodities and overproduction of capital (over-accumulation).

Likewise, the question to know whether the crisis 'centres' on the sphere of production or the sphere of circulation is largely meaningless. The crisis is a disturbance (interruption) of the process of enlarged reproduction; and according to Marx, the process of reproduction is precisely a (contradictory) unity of production and circulation. For capitalists, both individually (as separate firms) and as the sum total of firms, it is irrelevant whether more surplus-value has actually been produced in the process of production, if that surplus-value cannot be totally realized in the process of circulation. Contrary to many economists, academic and Marxist alike, Marx explicitly rejected any Say-like illusion that production more or less automatically finds its own market.

It is correct that in the last analysis, capitalist crises of overproduction result from a downslide of the average rate of profit. But this does not represent a variant of the 'monocausal' explanation of crisis. It means that, under capitalism, the fluctuations of the average rate of profit are in a sense the seismograph of what happens in the system as a whole. So that formula just refers back to the sum-total of partially independent variables, whose interplay causes the fluctuations of the average rate of profit.

Capitalist growth is always disproportionate growth, i.e. growth with increasing disequilibrium, both between different departments of output (Marx basically distinguishes department I, producing means of production, and department II, producing means of consumption; other authors add a department III producing non-reproductive goods – luxury goods and arms – to that list), between different branches and between

production and final consumption. In fact, 'equilibrium' under capitalism is but a conceptual hypothesis practically never attained in real life, except as a border case. The above mentioned tendency of 'overshooting' is only an illustration of that more general phenomenon. So 'average' capital accumulation leads to overaccumulation which leads to the crisis and to a prolonged phenomenon of 'underinvestment' during the depression. Output is then consistently inferior to current demand, which spurs on capital accumulation, first to a 'normal' level and then to renewed overaccumulation, all the more so as each successive phase of economic revival starts with new machinery of a higher technological level (leading to a higher average productivity of labour, and to a bigger and bigger mountain of produced commodities. Indeed, the very duration of the business cycle (in average 7.5 years for the last 160 years) seemed for Marx determined by the 'moral' life-time of fixed capital, i.e. the duration of the reproduction cycle (in value terms, not in possible physical survival) of machinery.

The ups and downs of the rate of the profit during the business cycle do not reflect only the gyrations of the output/disposable income relation; or of the 'organic composition of capital'. They also express the varying correlation of forces between the major contending classes of bourgeois society, in the first place the short-term fluctuations of the rate of surplus-value reflecting major victories or defeats of the working class in trying to uplift or defend its standard of living and its working conditions. Technological progress and labour organization 'rationalizations' are capital's weapons for neutralizing the effects of these fluctuations on the average rate of profit and on the rate of capital accumulation.

In general, Marx rejected any idea that the working class (or the unions) 'cause' the crisis by 'excessive wage demands'. He would recognize that under conditions of overheating and 'full employment', real wages generally increase, but the rate of surplus-value can simultaneously increase too. It can, however, not increase in the same proportion as the organic composition of capital. Hence the decline of the average rate of profit. Hence the crisis.

But if real wages do not increase in time of boom, and as they unavoidably decrease in times of depression, the average level of wages during the cycle in its totality would be such as to cause even larger overproduction of wage goods, which would induce an even stronger collapse of investment at the height of the cycle, and in no way help to avoid the crisis.

Marx energetically rejected any idea that capitalist production, while it appears as 'production for production's sake', can really emancipate itself from dependence on 'final consumption' (as alleged e.g. by Tugan-Baranowsky). While capitalist technology implies indeed a more and more 'roundabout-way-of-production', and a relative shift of resources from department II to department I (that is what the 'growing organic composition of capital' really means, after all), it can never develop the productive capacity of department I without developing in the medium and long-term the productivity capacity of department II too, admittedly at a slower pace and in a lesser proportion. So any medium or long-term contraction of final consumption, or final consumers' purchasing power, increases instead of eliminates the causes of the crisis.

Marx visualized the business cycle as intimately intertwined with a *credit cycle,* which can acquire a *relative* autonomy in relation to what occurs in production properly speaking (Marx, *g*, pp. 570 – 73). An (over)expansion of credit can enable the capitalist system to sell temporarily more goods that the sum of real incomes created in current production plus past savings could buy. Likewise, credit (over) expansion can enable them to invest temporarily more capital than really accumulated surplus-value (plus depreciation allowances and recovered value of raw materials) would have enabled them to invest (the first part of the formula refers to net investments; the second to gross investment).

But all this is only true temporarily. In the longer run, debts must be paid; and they are not automatically paid through the results of expanded output

and income made possible by credit expansion. Hence the risk of a *krach,* of a credit or banking crisis, adding fuel to the mass of explosives which cause the crisis of overproduction.

Does Marx's theory of crisis imply a theory of an inevitable final collapse of capitalism through purely economic mechanisms? A controversy has raged around this issue, called the 'collapse' or 'breakdown' controversy. Marx's own remarks on the matter are supposed to be enigmatic. They are essentially contained in the famous chapter 32 of volume I of *Capital* entitled 'The historical tendency of capitalist accumulation', a section culminating in the battle cry: 'The expropriators are expropriated' (Marx, *e*, p. 929). But the relevant paragraphs of that chapter describe in a clearly non-enigmatic way, an interplay of 'objective' and 'subjective' transformations to bring about a downfall of capitalism, and not a purely economic process. They list among the causes of the overthrow of capitalism not only economic crisis and growing centralization of capital, but also the growth of exploitation of the workers and of their indignation and revolt in the face of that exploitation, as well as the growing level of skill, organization and unity of the working class. Beyond these general remarks, Marx, however, does not go.

Marx and Engels on the Economy of Post-Capitalist Societies

Marx was disinclined to comment at length about how a socialist or communist economy would operate. He thought such comments to be essentially speculative. Nevertheless, in his major works, especially the *Grundrisse* and *Das Kapital,* there are some sparse comments on the subject. Marx returns to them at greater length in two works he was to write in the final part of his life, his comments on the *Gotha Programme* of united German social-democracy (Marx, *j*), and the chapters on economics and socialism he wrote or collaborated with for Engels' *Anti-Dühring* (1878). Generally his comments, limited and sketchy as they are, can be summarized in the following points.

Socialism is an economic system based upon conscious planning of production by associated producers (nowhere does Marx say: by the state), made possible by the abolition of private property of the means of production. As soon as that private property is completely abolished, goods produced cease to be commodities. Value and exchange value disappear. Production becomes production for use, for the satisfaction of needs, determined by conscious choice (*ex ante* decisions) of the mass of the associated producers themselves. But overall economic organization in a postcapitalist society will pass through two stages.

In the first stage, generally called 'socialism', there will be relative scarcity of a number of consumer goods (and services), making it necessary to measure exactly distribution based on the actual labour inputs of each individual (Marx nowhere refers to different quantities and *qualities* of labour; Engels explicitly *rejects* the idea that an architect, because he has more skill, should consume more than a manual labourer). Likewise, there will still be the need to use incentives for getting people to work in general. This will be based upon strict equality of access for all trades and professions to consumption. But as human needs are unequal, that formal equality masks the survival of real inequality.

In a second phase, generally called 'communism', there will be plenty, i.e. output will reach a saturation point of needs covered by material goods. Under these circumstances, any form of precise measurement of consumption (distribution) will wither away. The principle of full needs satisfaction covering all different needs of *different* individuals will prevail. No incentive will be needed any more to induce people to work. 'Labour' will have transformed itself into meaningful many-fold activity, making possible all-round development of each individual's human personality. The division of labour between manual and intellectual labour, the separation of town and countryside, will wither away. Humankind will be organized into a free federation of producers' and consumers' communes.

Selected Works

There is still no complete edition of all of Marx's and Engels's writings. The standard German and Russian editions by the Moscow and East Berlin Institutes for Marxism-Leninism, generally referred to as *Marx-Engels-Werke* (MEW), do not include hundreds of pages printed elsewhere (e.g. Marx's *Enthüllungen zur Geschichte der Diplomatie im 18. Jahrhundert* [Revelations on the History of 18th- century Diplomacy]), and several thousand pages of manuscripts not yet printed at the time these editions were published. At present, a monumental edition called *Marx-Engels-Gesamtausgabe* (MEGA) has been started, again both in German and in Russian, by the same Institutes. It already encompasses many of the unpublished manuscripts referred to above, in the first place a previously unknown economic work which makes a bridge between the *Grundrisse* and vol. 1 of *Capital*, and which was written in the years 1861–3 (published under the title *Zur Kritik der Politischen Oekonomie – Contribution to a Critique of Political Economy 1861–1863* in MEGA II/3/1–6, Berlin Dietz Verlag, 1976–1982). Whether it will include all of Marx's and Engels's writings remains to be seen.

In English, key works by Marx and Engels have been systematically published by Progress Publishers, Moscow, and Lawrence & Wishart, London; but this undertaking is by no means an approximation of the *Marx-Engels-Werke* mentioned above. The quality of the translation is often poor. The translations of Marx's and Engels's writings published by Penguin Books in the *Marx Pelican Library* are quite superior to it. We therefore systematically refer to the latter edition whenever there is a choice. Marx's and Engels's books and pamphlets referred to in the present text are mostly in chronological order:

(Marx *a*) *Die Deutsche Ideologie* (1846), together with Friedrich Engels.

(Marx *b*) *Manifest der Kommunistischen Partei* (1848), written in collaboration with Friedrich Engels. In English: *Manifesto of the Communist Party,* in *Marx: The Revolutions of 1848,* Harmondsworth: Penguin Books, 1973.

(Marx c) Zur Kritik der Politischen Oekonomie (1858). In English: Contribution to the Critique of Political Economy, London: Lawrence & Wishart, 1970.

(Marx *d*) *Grundrisse der Kritik der Politischen Oekonomie* (written in 1858-1859, first published in 1939). English edition: *Foundations of a Critique of Political Economy,* Harmondsworth: Penguin Books, 1972.

(Marx *e*): *Das Kapital, Band I* (1867). In English: *Capital,* Vol. I, Harmondsworth: Penguin Books, 1976.

(Marx *f*) *Das Kapital, Band II,* published by Engels in 1885. In English: *Capital,* Vol. II, Harmondsworth: Penguin Books, 1978.

(Marx *g*) *Das Kapital, Band III,* published by Engels in 1894. In English: *Capital,* Vol. III, Harmondsworth: Penguin Books, 1981.

(Marx *h*) *Theorien über den Mehrwert*, published by Karl Kautsky 1905–10. In English: *Theories of Surplus Value,* Moscow: Progress Publishers, 1963.

(Marx *i*) *Lohn, Preis und Profit,* written in 1865. In English: *Wages, Price and Profits,* in *Marx-Engels Selected Works,* Vol. II, Moscow: Progress Publishers, 1969.

(Marx j) Kritik des Gothaer Programms, written in 1878 in collaboration with Engels. In English: Critique of the Gotha Programme, in Marx-Engels: The First International and After, Harmondsworth: Penguin Books, 1974.

(Marx *k*) *Resultate des unmittelbaren Produktionsprozesses* (unpublished section VII of Vol. I of *Capital*), first published in 1933. In English: *Results of the Immediate Process of Production, Appendix to Capital*, Vol. I, Harmondsworth: Penguin Books, 1976.

(Marx *l*) *Marx-Engels: Briefwechsel (Letters)*. There is no complete English edition of the letters. Some are included in the *Selected Works* in 3 vols, published by Progress Publishers, Moscow.

(Engels): *Anti-Dühring* (1878). The chapter on economy was written by Marx, who also read all the other parts and collaborated in their final draft. In English: *Anti-Dühring*, London: Lawrence & Wishart, 1955.

Bibliography

There are innumerable books and articles devoted to comments or elaborations on Marx's economic thought, or which criticize them. We list here those works we refer to in the above text, as well as those we consider the most important ones (based, needless to say, upon subjective judgement).

Baran, P.A. 1957. *The political economy of growth*. London: John Calder.

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

Böhm-Bawerk, E. von. 1896. *Zum Abschluss des Marxschen Systems*. English edn, *Karl Marx and the Close of this System*, including a reply by Rudolf Hilferding, *Böhm-Bawerk's Criticism of Karl Marx*, ed. P.M. Sweezy. New York: Augustus M. Kelley, 1949.

Bortkiewicz, L. von. 1907. *Zur Berichtigung der grundlegenden theoretischen Konstruktion von Marx im Dritten Band des 'Kapital'.* Trans. P.M. Sweezy as 'On the Correction of Marx's Fundamental Theoretical Construction in the Third Volume of *Capital'.* In Böhm-Bawerk (1949), 197–221.

Braverman, H. 1974. *Labor and monopoly capital: The degradation of work in the twentieth century.* New York/London: Monthly Review Press.

Bronfenbrenner, M. 1970. *The vicissitudes of Marxian economics*. London.

Bukharin, N. 1914. *Imperialism and world economy.* English trans. London: M. Lawrence, 1915.

Bukharin, N. 1926. *Imperialism and the accumulation of capital.* Trans. R. Wichmann, ed. K.J. Tarbuck. London: Allen Lane, 1972.

Dobb, M. 1937. *Political economy and capitalism: Some essays in economic tradition.* London: G. Routledge & Sons. Reprinted Westport, Conn.: Greenwood Press, 1972.

Emmanuel, A. 1969. *L'échange inégal. Essai sur les antagonismes dans les rapports économiques internationaux*. Paris: François Maspero. Trans. B. Pearce as Unequal Exchange: A Study of the Imperialism of Trade. London: New Left Books, 1972.

Grossman, H. 1929. *Das Akkumulations- und Zusammenbruchsgesetz des kapitalistischen Systems*. Leipzig: C.L. Hirschfeld.

Hayek, F.A. von. 1944. *The road to serfdom.* Chicago/London: University of Chicago Press/G. Routledge & Sons.

Hilferding, R. 1910. *Das Finanzkapital.* Vienna: Wiener Volksbuchhandlung. Trans. M. Watnick and S. Gordon, ed. T. Bottomore as *Finance Capital.* London: Routledge & Kegan Paul, 1981.

Itoh, M. 1980. *Value and crisis: Essays on Marxian economics in Japan*. London: Pluto.

Kolakowski, L. 1976–8. *Main currents of Marxism: Its rise, growth and dissolution,* 3 vols. Trans. P.S. Falla. Oxford: Clarendon Press, 1978.

Lange, O. 1963. *Political economy,* 2 vols. Trans. and ed. P.F. Knightsfield. Oxford: Pergamon Press; Warsaw: PWN – Polish Scientific Publishers.

Lange, O. and F.M. Taylor. 1938. *On the economic theory of socialism,* 2 vols. Minneapolis: University of Minnesota Press.

Lenin, V.I. 1917. *Imperialism, last stage of capitalism*. Petrograd. English trans., *Imperialism, the Highest Stage of Capitalism.* Moscow: Foreign Languages Publishing House, 1947.

Leontief, W. 1938. The significance of Marxian economics for present-day economic theory. *American Economic Review,* Supplement, March. Reprinted in *Marx and modern economists,* ed. D. Horowitz. London: MacGibbon & Kee, 1968.

Luxemburg, R. 1913. *Akkumulation des Kapitals.* Trans. A. Schwarzschild, with an introduction by J. Robinson, as *The Accumulation of Capital.* London: Routledge & Kegan Paul, 1951.

Mandel, E. 1962. *Traité d'économie marxiste.* Paris: R. Juillard. Trans. B. Pearce as *Marxist Economic Theory*, 2 vols. London: Merlin Press, 1968.

Mandel, E. 1972. *Der Spätkapitalismus.* Frankfurt am Main: Suhrkampf. Trans. J. De Bres as *Late Capitalism.* London: New Left Books. Revised edn, 1975.

Mandel, E. 1980. *Long waves of capitalist development: The Marxist interpretation*. Cambridge: Cambridge University Press.

Mandel, E., and A. Freeman, eds. 1984. *Ricardo, Marx, Sraffa*. London: Verso.

Mattick, P. 1969. *Marx and Keynes: The limits of the mixed economy.* Boston: P. Sargent.

Mises, L. von. 1920. Die Wirtschaftsrechnung im sozialistischen Gemeinwesen. Trans. S. Adler as 'Economic Calculations in the Socialist Commonwealth'. In *Collectivist economic planning*, ed. F.A. von Hayek. London: G. Routledge & Sons, 1935.

Morishima, M. 1973. *Marx's economics: A dual theory of value and growth*. Cambridge: Cambridge University Press.

Nutzinger, H.G., and E. Wolfstetter, eds. 1974. *Die Marx'sche Theorie und ihre Kritik,* 2 vols. Frankfurt am Main: Campus.

Pareto, V. 1966. *Marxisme et économie pure*. Geneva: Droz.

Popper, K. 1945. *The open society and its enemies,* 2 vols. London: G. Routledge & Sons.

Robinson, J. 1942. *An essay on Marxian Economics*, 2nd edn. London: Macmillan, 1966.

Rosdolsky, R. 1968. *Entstehungsgeschichte des Marxschen 'Kapital'*. Frankfurt am Main: Europäische Verlagsanstalt. Trans. P. Burgess as *The Making of Marx's Capital*. London: Pluto Press, 1977.

Rubin, I.I. 1928. *Essays on Marx's theory of value.* Moscow. English trans., Detroit: Black and Red, 1972.

Schumpeter, J. 1942. *Capitalism, socialism and democracy.* New York/London: Harper & Brothers.

Sraffa, D. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.
Sternberg, F. 1926. *Der Imperialismus*. Berlin: Malik-Verlag.
Sweezy, P.M. 1942. *The theory of capitalist development*. New York: Monthly Review Press.
Tugan-Baranowsky, M. von 1905. *Theoretische Grundlagen der Marxismus*. Leipzig: Duncker & Humblot.
Wygodsky, S. 1972. *Der gegenwärtige Kapitalismus*. Trans. C.S.V. Salt as *The Story of a Great Discovery: How Karl Marx wrote 'Capital'*. Tunbridge Wells: Abacus Press.

Bibliographic Addendum

Foley, D. 1986. *Understanding capital: Marx's economic theory*. Cambridge, MA: Harvard University Press.
Roemer, J. 1981. *Analytical foundations of Marxian economic theory*. Cambridge: Cambridge University Press.
Roemer, J. 1988. *Free to Lose: An introduction to Marxist economic philosophy*. London: Radius.
Outside of economics, important studies include:
Buchanan, A. 1982. *Marx and justice*. Totowa, NJ: Rowan and Allanhead.
Carver, T. 1998. *The post-modern Marx*. Manchester: Manchester University Press.
Cohen, G.A. 1988. *History, labour, and freedom: Themes from Marx*. Oxford: Oxford University Press.
Cohen, G.A. 2000. *Karl Marx's theory of history: A defence,* 2nd edn. Oxford: Clarendon Press.
Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.
Kain, P. 1989. *Marx and ethics*. Oxford: Oxford University Press.
Roemer, J., ed. 1986. *Analytical Marxism*. Cambridge: Cambridge University Press.
Rockmore, T. 2002. *Marx After Marxism*. Oxford: Basil Blackwell.
Van den Berg, A. 1989. *The immanent Utopia: From Marxism on the state to the state of Marxism*. Princeton: Princeton University Press.
Wolff, J. 2002. *Why read Marx today?* Oxford: Oxford University Press.
Wood, A. 2004. *Karl Marx,* 2nd edn. London: Routledge.
All of these are complemented by what remains a classic short introduction to Marx's ideas:
Berlin, I. 1985. *Karl Marx,* 4th edn. Oxford: Oxford University Press.
as well as the standard biography of Marx:
McLellan, D. 1996. *Karl Marx: A biography*, 3rd revised edn. London: Palgrave Macmillan.
Finally,
Bottomore, T., ed. 1983. *A dictionary of Marxist thought*. Cambridge, MA: Harvard University Press.
is helpful both as a survey and as a mechanism for overcoming language differences between Marxist approaches and others.

Marx's Analysis of Capitalist Production

Duncan Foley and Gérard Duménil

Abstract

This article discusses Marx's analysis of capitalism, including the concepts of historical materialism, class society, exploitation, commodity, value, money, capital, labour-power, value of labour-power, surplus-value, constant and variable capital, commodity law of exchange, capitalist law of exchange, equalization of the profit rate, prices of production, absolute and relative surplus value, the circuit of capital, simple and expanded reproduction, capital accumulation, centralization and concentration of capital, technical change, reserve armies of labour, rent, interest, commercial and bank profit, the falling rate of profit, viable technical change, and cyclical crises.

Keywords

Absolute rent; Abstract labour time; Biased technical change; Business cycles; Capital accumulation; Capitalism; Capitalist law of exchange; Circulating and fixed capital; Circulation of capital; Class; Commercial, industrial and money-dealing capital; Commodity; Commodity law of exchange; Commodity money; Competition; Composition of capital; Concentration and centralization of capital; Constant and variable capital; Crises of overproduction; Differential rent; Division of labour; Exchange value and use value; Exploitation; Fictitious capital; Fixed capital; Historical materialism; Increasing returns; Industrial capital; Innovation; Iron law of wages; Labour power; Labour theory of value; Law of value; Marx's analysis of capitalist production; Marxian transformation problem; Money; Money-dealing capital; Monopoly; Organic composition of capital; Over-accumulation of capital; Population

growth; Primitive accumulation; Private property; Productive and unproductive labour; Profit rate; Rate of exploitation; Rate of surplus value; Ricardo, D.; Say's Law; Slavery; Smith, A.; Socialism; Surplus; Surplus value; Technical change; Transformation problem; Use value; Value; Variable capital; Velocity of circulation

JEL Classifications
B1; B2; B5

Karl Marx's analysis of capitalist production is best understood in the context of his broad theory of human societies and their history, namely, *historical materialism.* This theory argues that, after passing through various stages in which societies are divided into classes and the exploitation of a majority of producers by a privileged minority prevails, humanity will finally eliminate classes and class domination by a revolutionary process conducted by the organized *proletariat* in capitalism. This revolutionary stand was based on a 'scientific' investigation of history in general and capitalism in particular, with a special emphasis on economics, always with a political perspective. Whether historical materialism has a scientific or ideological character obviously remains controversial between Marxists and non-Marxists: Marxist theory is considered a discredited doctrine of the past by non-Marxists, while Marxists consider mainstream social and economic thinking as a continuing apologetics of capitalism.

After an introductory section devoted to locating the capitalist mode of production as a particular epoch in human history, the main focus below is on Marx's analysis of capitalist production. There are two facets to the theory of capital in the strict sense: *surplus value* (exploitation), and *the circuit of capital* (its 'circulation'). These are introduced separately, and then gradually combined in the analysis of more complex phenomena. Finally, we consider three broad sets of basic mechanisms directly related to the hold of capital on the functioning of the economy: (1) competition, (2) accumulation, technological and distributional changes, and (3) crises and the business cycle. We do not consider other important aspects of Marx's thinking such as his analysis of class struggle, and his theory of the state. The interpretation of even very fundamental aspects of Marx's thought remains contested among Marxist scholars. The bibliography contains a selective list of works that represent some of these different perspectives.

The Capitalist Mode of Production

The historical materialist point of view starts from the observation that all human societies must produce in order to reproduce both individuals and society itself. Production in this general sense always involves the combination of human labour with previously produced means of production and the natural resources of the earth. With the emergence of settled agriculture a *surplus product* over and above what is necessary for reproduction becomes possible. In societies with a surplus product, class exploitation, an institutionalized form of inequality, arises. Societies divide into a small *exploiting class* which appropriates, controls, and distributes the surplus product created by the labour of a much larger *exploited class* of producers who receive on average only what is necessary for their reproduction. Marx and Engels distinguish two aspects of these *class societies.* The *forces of production* comprise the population, natural resources, and technology which make a surplus product possible; the *social* relations *of production* comprise the institutional framework (such as property relations) through which the exploiting class appropriates the surplus product. The forces and social relations of production together constitute a *mode of production.* For example, in the slave mode of production characteristic of ancient Greek and Roman civilizations, the institution of slavery sustained by military force and political power was the means through which slave-owners appropriated a surplus product created by the labour of slaves, who received a minimum

subsistence. In the feudal mode of production, the institutions of serfdom sustained by military force and religious and political power were the means through which the lords of the manor appropriated a fraction of the labour time of serfs, who also laboured in their own fields to feed and reproduce themselves (or the serf had to pay a rent in kind or, later, in money, in addition to various taxes). This is what *exploitation* means in Marx's thought: to live on the product of the labour of other people.

From the historical materialist point of view, capitalism is a class society in which the institutions of private property in the means of production and free wage labour are the means through which capitalists appropriate the *surplus value* created by workers producing commodities (or services), who receive wages. In feudalism, the exploitation of the serfs was transparent: the serfs worked a certain part of the week on their own plots for their own subsistence, and a certain part of the week on the lord's land to supply his consumption and armies. Marx's theory of capitalism demonstrates that, though the mechanism of capitalist exploitation through the social relation of wage labour based on the formal legal equality of workers and employers is less transparent, capitalists also appropriate the surplus labour time of the workers. Capitalism, therefore, defines a specific stage of the history of class societies. Capitalism's decentralized, highly competitive organization creates powerful incentives for the rapid development of the forces of production through population growth, technical innovation, and a widening division of labour, but it is unable to control the forces it has itself stimulated.

Marx and Engels expected that the capitalist working class (the *proletariat*), once it had a clear understanding of capitalist exploitation and reached a high degree of organization, would overthrow the social relations of capitalism in a revolution to establish a classless society based on social control of the large surplus product made possible by the forces of production developed by capitalism. A violent transition was required, the *dictatorship of the proletariat,* to attain socialism and finally communism, marking the end of the 'prehistory' of humanity. Marx developed this analysis in collaboration with Friedrich Engels in *The Communist Manifesto* (1848).

Marx's main work, *Capital,* is devoted to the analysis of capitalist production. The first volume was published, in 1867, while Marx (1818–83) was still alive. Volumes II and III were published later by Friedrich Engels, from extensive notebooks still in draft form at the time of Marx's death. In what follows, we refer to *Capital* by volumes and chapters; for example, 'III, 25' means Chapter 25 of Volume III. References and quotations can be found on Internet, for example in the *Marx/ Engels Library,* http://www.marxists.org/archive/marx/works, or in Marx (1976, 1978, 1981). We have put square brackets around our own interpolations in quotes; everything else comes from the source.

The Definition of Capital (I, 4)

Marx defines capital as *value* (to be defined below) participating in a dynamic process of self-expansion. A capitalist spends money to hire workers and buy means of production, and then sells the resulting output for enough money to cover his initial outlay and secure a profit (the form taken by 'surplus value'). In this process value appears in various forms: first under the form of money; then as the value of productive inputs (labour power, raw materials, machinery, and buildings); then as the value of the commodities produced; and finally as money value again after the produced commodities have been sold. This process of capital is pointless unless, as is normally the case when capitalists make a profit, the money realized in the sale of commodities is greater than the money initially spent to start the process. Capital is not value as such, but value in movement:

> If we pin down the specific forms of appearance assumed in turn by self-valorising value, in the course of its life, we reach the following elucidation: capital is money, capital is commodities. In truth, however, value is here the subject of a process in which, while constantly assuming the form in turn of money and commodities, it changes its own magnitude, throwing off surplus- value from itself considered as original value. (I, 4)

Two aspects of capital are present in this definition: (1) capital is expanding value; and (2) capital value changes its form. These two aspects of capital are also called the *process of self-expansion* (sometimes called *valorization*), and the *process of circulation of capital* (or circuit of capital). Marx means here that: (1) the capitalist invests a certain capital with the intent of making profits (expansion); (2) capital is invested in commodities and money, and constantly passes from one form to the other (for example, when an output is sold, value changes form from commodity to money).

The first two volumes of *Capital* treat the processes of *self-expansion* and *circulation* of capital separately (with a few exceptions); the third volume considers the combination of these two elements. Before entering into the analysis of capital, it is necessary, however, to introduce two other preliminary concepts, *commodity* and *money,* and the related concepts of *value* (at the centre of the definition of capital) and *price*, to which Marx devotes the first three chapters of Volume I, prior to the analysis of capital. In Volumes I and II, the three concepts are considered successively: commodity (including value), money (including price), and capital (valorization and circulation). (This outline is logical, not historical: historically commodities and money reach their full development only with the capitalist mode of production.) We will follow this outline in our exposition here.

Commodities, Value, Money, and Prices

Commodities and Value (I, 1)

A *product* is the result of human labour, working with produced means of production and the natural resources of the earth. Useful products become *commodities* when they are regularly exchanged rather than being consumed directly by their producers. 'Useful' must be taken in a very broad sense as something desired by someone, for whatever reason. A producer who exchanges his product receives social recognition for his own labour in the form of the other commodities he acquires. Marx denotes the labour time required for the production of a commodity under average conditions, as *socially necessary labour time.* As the outcome of a parcel of social labour time, the commodity has an *exchange value,* or more briefly a *value.* Thus, according to Marx (who here follows Adam Smith), a commodity has a dual character as: (1) *object of utility,* or equivalently a *use value,* and (2) an *exchange value,* or *value.* The value of the commodity is the sum of labour embodied in previously produced inputs, *dead labour*, and newly incorporated labour, *living labour.* Marx sometimes calls this definition the *law of value,* although he rarely uses the expression. Later economists often refer to this framework as the *labour theory of value.*

The dual character of the commodity is reflected on labour itself. The concrete quality of labour (weaving, computer-programming) corresponds to the use-value aspect of the commodity it produces. But all categories of social labour materialized in the production of commodities have in common the ability to produce exchange values and, as such, are defined as *abstract labour.* There is no a priori rule accounting for this process of abstraction. Exchange dissolves the specific character of concrete labours, and the repetition of exchange establishes their quantitative equivalence. If one category of concrete labour is not adequately compensated, its supply will decline, and its wage will rise. In a similar manner, it is exchange which establishes the normal degree of intensity, skill, and technical efficiency in production.

Abstracting from the capitalist character of production, commodities would 'normally' exchange in proportion to their values. For example, if the value of commodity A is twice that of commodity B, one unit of A will exchange for two of B. If the exchange ratio were only one B for one A, producers of A would switch to producing B; a shortage of A would ensue and the exchange value of A would rise. This is the *commodity law of exchange,* sometimes confused with the law of value. The distinction is important because the law of value is a fundamental characteristic of commodity production, whether commodities exchange in proportion to their values or not. (In competitive capitalist economies they typically do not, as we will see.)

Money and Prices (I, 3)

We begin with the definition of money, and its first function as *measure of value,* and introduce the other functions of money, and the concept of the *price form* of value.

The value of commodities cannot be expressed on the market directly in abstract labour time (which nobody can observe or measure). In the exchange of two commodities, such as linen for a coat, the value of one commodity is expressed in the body of the other (measured in units such as a length or weight) as its direct *equivalent.* With the repetition of exchange, some specific commodity, such as gold, will emerge as a *socially accepted general equivalent,* that is, as *money.* Thus for Marx the original function of money is as *measure of value.* In addition to its function as measure of value, money comes to serve as *medium of circulation* if purchases and sales are paid for directly, and as *means of payment* if payment is deferred. Value can be accumulated temporarily in money hoards. Another function of money is, therefore, as a *store of value* (though any durable, valuable commodity can serve as a store of value).

Prices are values as expressed in monetary units. They are *forms of value.*

When commodities exchange at prices proportional to their values, the price of a commodity expresses the socially necessary (abstract) labour time required for its production of this commodity, qualitatively and quantitatively in a straightforward manner. This is the framework of Volumes I and II. But the prices of commodities may deviate from their values, and we will later return to this issue. The State can establish a *standard of price* by defining a local currency unit such as the franc or dollar as a certain amount of gold or other money commodity. Valueless tokens, 'symbols or tokens of value' in Marx's words, such as paper currency, may also be circulated in place of commodity money:

> In the same way as the exchange-value of commodities is crystallised into gold money as a result of exchange, so gold money in circulation is sublimated into its own symbol, first in the shape of worn gold coin, then in the shape of subsidiary metal coin, and finally in the shape of worthless counters, scraps of paper, mere *tokens of value.* (Marx 1859, 2.B.2.c)

Money also takes the form of a stock of purchasing power in an account in a financial institution. In contemporary capitalism, there is no commodity money.

The Monetary Expression of Value and the Quantity of Money

Inherent in Marx's theory is the relation between abstract labour time and its price form in money terms. There is a quantitative aspect to this relation. The ratio – for example, dollars per hour of abstract socially necessary labour time – can be called the monetary expression of labour time, or the monetary expression of value.

The determination of this ratio, which is a way of looking at the general price level in an economy, is discussed by Marx in his critique of Ricardo's quantity of money theory of prices, under the assumption of the existence of a commodity money. Marx explains that the quantity of money required to circulate the mass of commodities produced in any period depends on the quantity of the commodities exchanged, their money prices, determined by their costs of production, and the velocity of money, the average number of transactions in which each unit of money participates in the period (an institutional characteristic). Money flows in and out of hoards (reserves) to accommodate the requirements of circulation. He interprets this principle as governing the quantity of money required for purchases and sales, in contrast to Ricardo's quantity of money theory of prices, which sees the prices of commodities adjusting to a given quantity of money. In Marx's theory the general level of prices is determined by the relative costs of production of the money commodity and other commodities when a commodity like gold is used as money. (The critique of Ricardo's theory is developed in Marx 1859, 2.C.)

The Theory of Surplus Value

The labour theory of value is the foundation of Marx's theory of exploitation, or surplus value.

When a produced commodity is purchased or sold no new value is created. If a commodity sells at a price proportional to its value, given the monetary expression of value, the buyer and seller exchange money and commodity representing equal values. If the commodity sells above or below its value, the value gained by one party is just offset by the value lost by the other.

Productive Labour-Power and Surplus-Value (I, 7–9)

Marx explains surplus value in relation to the purchase of the labour power of waged workers. The capability to work, denoted as *labour power*, is a commodity, with a use-value and a value. The *use value* of labour power is labour itself, because a capitalist buys labour power to obtain the right to use the labour of the worker. The *value* of labour power is the value equivalent of the purchasing power of the wage on the commodities the worker can buy. (We will discuss later Marx's view of the actual purchasing power of workers.)

Only 'productive workers', that is, workers involved directly in production within capitalist enterprises, produce new value in Marx's analysis, in contrast to 'unproductive workers', whose labour power is employed by capitalists to maximize their profit rate. If the value of the labour power of productive workers is less than the value they produce, capitalist production on average adds more value in the production of commodities than it expends in hiring workers. (One can, equivalently, say that the money wage must be smaller than the monetary expression of the labour time expended by the average worker.) Because capitalist production can produce a surplus over the subsistence of productive workers, typically the value of labour power is smaller than the value labour produces, and surplus value results.

Thus, labour power has a property not shared by other commodities. While the purchase and sale of a produced commodity can only redistribute a given value between buyer and seller, the capitalist's purchase and use of labour-power, in contrast, results in the creation of surplus-value. The capitalist buys labour-power at a wage reflecting the necessary labour time required by the production of the consumption basket of the worker, say, four hours a day, but on average the worker can work longer, say, eight hours. Thus, the capitalist can appropriate *surplus labour,* here four hours, in the form of surplus value. (If the monetary expression of labour time is ten dollars per hour, the surplus value created by an average worker in a day under these assumptions would be 40 dollars.) Under the wage system, once a capitalist has paid a worker the agreed wage, the product of the worker's labour and its value belong to the capitalist. The production of surplus value is thus compatible with transactions at prices proportional to values, including the purchase of labour power at a wage proportional to the value of productive labour power. Marx argues that capitalist exploitation does not violate the commodity law of exchange, that is, it would take place even if all commodities exchanged at prices proportional to their values.

The actual appearance of labour power available for hire historically depends on two preconditions. First, workers must be legally free to sell their labour power. This explains the historic hostility of capitalism to bound forms of labour such as serfdom and slavery. Second, workers cannot have access to their own means of production, such as the feudal commons, so that they have no choice but to sell their labour power to the owner of means of production to live. This explains the historic support of capitalism for the enclosure of common lands and their conversion into private property. Marx devotes the last part of Volume I to *primitive accumulation,* the actual historical process through which the capitalist mode of production came into being. There he shows how, in the first steps of accumulation in England, the availability of labour power was achieved by way of straightforward social violence. The *enclosure* of common lands deprived the rural population of its old conditions of reproduction, and subjected it to the dependency on capital. It is important to keep such mechanisms in mind in the investigation of the historical dynamics of capitalism. Marx emphasizes the crucial historical importance of the transformation of produced means of production and labour,

which are universal aspects of human production, into the specific commodity forms of capital, including labour power.

The value of the produced inputs the capitalist purchases to undertake production is recovered in the sales price unchanged, so that Marx calls it *constant capital*, denoted by the symbol *c*. The value of the labour power the capitalist buys as an input to production, on the other hand, is recovered in the sales price expanded by the addition of the surplus value, so that Marx calls it *variable capital*, denoted by the symbol *v*. The sum of constant capital, *c*, variable capital, *v*, and surplus value, *s*, is the total value of the product. The sum $c + v$ is the total cost of the commodity. The sum $v + s$ is the *living labour,* as opposed to *dead labour, c,* and measures the *value added* by the production process. The *rate of surplus value, s*/*v*, is the ratio of *unpaid* to *paid* labour time, so that Marx also calls it the *rate of exploitation.* The ratio *c*/*v*, which measures the ratio of dead to living labour in the cost of the commodity, is the *value composition* of capital.

This decomposition of the value of a commodity is parallel to the income statement of a capitalist firm, which exhibits *profit* (Marx's surplus-value, *s*) as the difference between sales price (Marx's value of the commodity, $c + v + s$), and the cost of the means of production and wages required to produce the commodity (Marx's $c + v$).

Absolute and Relative Surplus Value, Manufacture and Industry (I, 12–16)

Identifying surplus value as surplus labour time does not tell what determines its magnitude and variation. Many natural, social and political conditions are involved, and vary historically. Labour performed by members of the family at home, women in particular, crucially affects the level of exploitation compatible with the reproduction of the workers and their families. In his analysis of surplus value in Volume I, Marx introduces important developments concerning the historical transformation of technology and organization.

Surplus value can be increased in two analytically distinct ways (which can be combined in real production): first, by lengthening the duration of labour time without increasing the value of labour power, *absolute surplus value*; second, by diminishing the value of labour power by cheapening worker's consumption through productivity gains holding the duration of labour time constant, *relative surplus value.* In Marx's view relative surplus value is the origin of the most important developments in the historical transformation of the organization of labour and technology by capitalism.

Marx sees distinct periods in which this transformation of production took different forms. In 'manufacture', a large number of individual workers, each processing his or her own means of production, are brought together in one location primarily for the purpose of increasing the capitalist's surveillance and control of production (which Marx describes as the 'formal subsumption' of labour to capital). In 'large-scale industry', the capitalist takes the further step of imposing a detailed division of labour on the production process, transforming the workers' relation to the production process (which Marx describes as the 'real subsumption of labour to capital'). Both technology and organization enter into these transformations. In manufacture, workers originally worked with the same tools they previously used in production at home; in large-scale industry, by contrast, capital has completely transformed technology and the organization of labour.

We will return to Marx's theory of technical change in capitalism in the discussion of the falling rate of profit.

The Circulation of Capital

As defined earlier, capital is self-expanding value moving through various forms (money, commodity…). We now turn to the analysis of the circulation of capital. The emphasis is on the motion from one form to the other, and the coexistence of the various fractions of capital under the three forms at a given point in time.

The Circuit of Capital (II, 1–4)

A capitalist spends money to buy inputs (means of production and labour power); organizes

production; stockpiles and sells the resulting product; and realizes a certain amount of money in sales revenue, normally larger than the original capital outlay. Each atom of capital goes through the various *forms:* money-capital, M, commodity capital in the form of inputs to production, C, *productive capital, P,* the value of partially finished commodities and plant and equipment in the workshop, and again commodity capital in the form of inventories of commodities awaiting sale, C', and finally returning to money through the sale of the produced commodities, M'. Marx represents this sequence in a diagram of the *circuit of capital*:

$$M — C \ldots P \ldots C' — M'$$

Here M is the money the capitalist uses to buy inputs to production C, P represents the actual production process, and C' are the produced commodities which are sold for money M'. The dashes represent purchase and sale of commodities on the market. The circuit is a chain, which can be viewed as beginning in M, C, or P, the *circuits of money, commodity,* and *productive capital,* three distinct formula of the same circuit.

The speeds at which the values of the various components of capital go through the productive form of capital, P, can be quite different. The value of some components, like raw materials, returns quickly to the money form in the sale of the commodity, while others like the value of buildings and machinery (whose value is only transferred to the product along their service life) returns only after a long period of time. From these differences in turnover time follows the distinction between *circulating* and *fixed capital*.

Capital is also a stock of value at any point in time. All the circuits overlap simultaneously: at the same moment new means of production and labour-power are being purchased while production is going on and finished output is being sold. The capital of a capitalist is the total value, tied up at any moment in these circuits. The total capital, K, is divided into three component stocks: *money capital, M, commodity capital, C,* and *productive capital, P.* The sum $K = M + C + P$ parallels the total of the assets on the capitalist's balance sheet.

Industrial, Commercial and Money-Dealing Capital (III, 16; III, 19)

Industrial capital undergoes the complete circuit of capital as above, taking on the forms M, C, and P in turn. Some capitals, however, are specialized to limited segments of the circuit. The first is *commercial capital*, which buys finished commodities from industrial capitalists to sell them to final purchasers, in the reduced circuit M—C—M': commercial capitalists buy in order to sell the same commodity. The second category, *money-dealing capital*, refers to the technical activity of banks in handling money payments into and out of accounts (and the exchange of currencies). Since no productive labour is expanded in these circuits, no surplus value is created. How industries engaged in such activities can make profits is part of the theory of competition considered below.

Marx's Schemes of Reproduction (II, 18–21)

Although Volume II of *Capital* is devoted to the circulation of capital, the analysis of the *schemes of reproduction,* combines valorization (c, v, s) and circulation (M, C and P).

Three departments are distinguished which produce the physical commodities to satisfy the demand emanating from c, v, and s: Department I produces means of production, Department II commodities consumed by workers, and Department III commodities consumed by capitalists. If all of the surplus value is consumed, no accumulation takes place, and the size of the capitalist economy remains unchanged, the case of *simple reproduction.* If a fraction of the surplus value is accumulated, the corresponding purchasing power is spent on additional means of production, and the capitalist economy expands, the case of *expanded reproduction.*

Marx assumes that all capital in the three industries accomplishes exactly one circuit: at the beginning and at the end of the period, all capital is assumed to be under the form C (the stocks of means of production and worker and capitalist consumption goods waiting to be sold). In this

setting reproduction requires certain proportionalities to hold: for example, in simple reproduction the value added of Department I must equal the constant capital of Departments II and III.

In this framework, Marx considers two types of issues. The first issue is the definition of output and its relation to income. The *net product* is the value of the final product, C', minus the value of what is now denoted as intermediate inputs, either produced in the previous period, in C, or during the present period but purchased as inputs by firms. Marx shows that the value of this net product is equal to total income or *value added*, as in contemporary national accounting, the sum of wages and surplus value (including rent, interest and profit as we will see): $v + s$. Second, Marx investigates the circulation of money. He attempts to demonstrate how the money thrown into circulation by capitalists returns as sales revenue, taking into account the activities of a sector producing the money commodity if such a money commodity exists.

The Functions of the Capitalist and Their Delegation to Employees (II, 6)

Being a capitalist is not a sinecure: both the appropriation of surplus value and the circuit of capital require active attention. In contemporary language: enterprises must be managed. Marx refers to these tasks as 'capitalist functions', in particular commercial transactions:

> The transformations of the forms of capital from commodities into money and from money into commodities are at the same time transactions of the capitalist, acts of purchase and sale. The time in which these transformations of forms take place constitutes subjectively, from the standpoint of the capitalist, the time of purchase and sale; . . . the time in which the capitalist buys and sells and scours the market is a necessary part of the time in which he functions as a capitalist, i.e., as personified capital. It is a part of his business hours. (II, 6)

The tasks considered are variegated, from the overseeing of labour in the workshop to the acceleration of the circuit of capital (as in the market activities mentioned above). All these tasks are unproductive, though they are useful. Their purpose is the *maximizing of the profit rate* of the capitalist. (The profit rate is defined below in the treatment of competition.)

The capitalist delegates some of these unproductive tasks to employees. They require means of production as well as labour power, like industrial capitalist production, though they produce no value. The wage and capital costs of these unproductive activities are a deduction from the surplus value. Marx denotes them as 'costs', in particular *costs of circulation* (the control and acceleration of the circuit of capital). As a consequence Marx categorizes some wage labour employed in capitalist production as unproductive, as in, for example, the case of overseers and employees in trade.

The Distribution of Surplus Value as Income

In Volume III, surplus value in its relation to both self-expansion and circulation is renamed *profit*. Profit is a form of surplus value. Once extracted, surplus value is at the origin of various categories of incomes, which appear as deductions from profit. The payment of such incomes to agents who employ no labour is thus consistent with the labour theories of value and surplus value. These channels of distribution of surplus value correspond to specific fractions of ruling classes in capitalism, such as active capitalists (entrepreneurs), money capitalists and landowners.

Interest and Profit of Enterprise. Interest-Bearing Capital (III, 21–3)

Some capitalists do not engage directly in capitalist production, but put their capital at the disposal of another functioning industrial capitalist, the *active capitalist* (or *entrepreneur*). This transaction may take the form of a loan in exchange for a share of the surplus value as *interest*, or the purchase of shares of stock in the firm which pays *dividends*. Marx treats both cases as *interest-bearing capital*, and this category of capitalists as *money capitalists* (sometimes referred to as 'financial capitalists'). Marx explains interest as a portion of the surplus-value realized by active capitalists. The profit remaining after the active capitalist has paid dividends and interest is *profit*

of enterprise. The existence of a developed loan market with a uniform rate of interest (for each maturity and risk of the loan) leads active capitalists to regard their own capital as loan capital, and to impute interest on it as an opportunity cost. Thus profit of enterprise appears as a kind of wage to the entrepreneurial activities of the active capitalist.

Rent (III, 38, 45)

Owners of scarce natural resources ('land' in the terminology of the classical political economists) also receive incomes in deduction from profits, in the form of rents. Due to their monopoly ownership of specific pieces of land, landowners can bargain with individual capitalists for a share of the surplus-value as rent (or royalties in other instances). How rents are quantitatively determined can only be examined in relation of the theory of competition.

Finance

Banking Capital and Money Capitalists (II, 19; III, 29)

The tasks of money-dealing capital are performed by banks. This represents their first source of income.

Banks also concentrate and use available masses of capital. One source of funds for banks is the idle balances of money in the economy, which are deposited in bank accounts. Thus, the money capital of enterprises is pooled within banks together with the balances of money held by other agents, such as households. While individual balances fluctuate, the aggregate pools are much more stable. A second source of funds is the capital of money capitalists (interest-bearing capital, including stock shares), who, instead of dealing directly with entrepreneurs, use banks as intermediaries. (Marx is aware of the capability of banks to 'create' money, but his view of banking mechanisms remains dominated by intermediation.) The theory of banking capital unites these two facets of the theory of capital: money-dealing capital and the handling of the capital of money capitalists.

Besides the management of accounts, the main function of banks is to make these funds available to agents seeking financing. Banks actually become the 'administrators' of the capital of money capitalists, and 'confront' capital as used by enterprises:

> Borrowing and lending money becomes their [banks'] particular business. They act as middlemen between the actual lender and the borrower of money capital. Generally speaking, this aspect of the banking business consists of concentrating large amounts of the loanable money capital in the bankers' hands, so that, in place of the individual money-lender, the bankers confront the industrial capitalists and commercial capitalists as representatives of all money-lenders. They become the general managers of money capital. On the other hand by borrowing for the entire world of commerce, they concentrate all the borrowers vis-à-vis all the lenders. A bank represents a centralisation of money capital of the lenders, on the one hand, and, on the other, a centralisation of the borrowers. (III, 25)

It is in these pages of Volume III of *Capital* that Marx analyses the issuance of paper currency by private banks and the Bank of England.

Fictitious Capital and Financial Instability (III, 25)

Marx's original definition of capital, as value in a movement of self-expansion, does not apply to securities like Treasury bills, or even to the stock shares of corporations. To refer to these securities, Marx uses the phrase *fictitious capital.* A public bond is in no way 'fictitious' for its holder, but it has no counterpart in the *M*, *C* and *P* of the circuit of capital. Once bonds or equities have been sold by a capitalist firm and are being traded on a secondary market, their values are also fictitious. The emergence of a market interest rate leads to the phenomenon of the *capitalization* of income flows such as the interest on government debt and dividends on equity: the market, where expectations concerning the future of these flows are taken into account, assigns a principal value to any flow of income. Thus, the accumulation of capital is paralleled in capitalism by that of such fictitious capital. Marx sees this capitalization of revenue flows as a source of instability.

The Institutional Framework of Modern Capitalism (III, 21–3)

As noted earlier, with the development of capitalism, the functions of the active capitalist are gradually delegated to managers and employees. This configuration, in which funding is provided by money capitalists with banks acting as intermediary, and the bulk of capitalist functions is delegated to a salaried personnel is that of modern capitalism:

> But since, on the one hand, the mere owner of capital, the money capitalist, has to face the functioning capitalist, while money capital itself assumes a social character with the advance of credit, being concentrated in banks and loaned out by them instead of its original owners, and since, on the other hand, the mere manager who has no title whatever to the capital, whether through borrowing it or otherwise, performs all the real functions pertaining to the functioning capitalist as such, only the functionary remains and the capitalist disappears as superfluous from the production process. (III, 23)

The Trinity Formula of Capital and Classes in Capitalism (III, 48; III, 52)

A major objective of *Capital* is to establish surplus value as the source of all incomes in capitalist society except wages. But capitalist practice hides this origin of capitalist incomes in what Marx calls the 'trinity formula':

> Capital—profit (profit of enterprise plus interest), land—ground-rent, labour—wages, this is the trinity formula which comprises all the secrets of the social production process. (III, 48)

Actually, this configuration is again altered in what we called above the institutions of modern capitalism:

> Furthermore, since as previously demonstrated interest appears as the specific characteristic product of capital and profit of enterprise on the contrary appears as wages independent of capital, the above trinity formula reduces itself more specifically to the following: Capital—*interest, land—ground-rent, labour—wages*, where profit, the specific characteristic form of surplus-value belonging to the capitalist mode of production, is fortunately eliminated. (III, 48)

To Marx, this trinity formula is 'irrational', because it confuses the source of incomes in the distribution of surplus-value with the role of necessary inputs in the production of use-values.

Volume III of *Capital* stops on a single-page chapter (obviously incomplete), entitled 'Classes'. There Marx establishes a straightforward relationship between his analysis of incomes and the fundamental class pattern of capitalism:

> The owners merely of labour-power, owners of capital, and land-owners, whose respective sources of income are wages, profit and ground-rent, in other words, wage-labourers, capitalists and land-owners, constitute the three big classes of modern society based upon the capitalist mode of production. (III, 52)

To this one could add fractions of capitalist classes corresponding to the various circuits of capital and the division of surplus value as above: (1) industrial capitalists, commercial capitalists, bankers, and (2) entrepreneurs (active capitalists) and money capitalists.

The Distribution of Surplus Value Through Competition

The analysis of capitalist production we have summarized so far, based on the idea that surplus value (and hence capitalist profit) arises from the exploitation of productive labour, runs counter to the apparent linkage of profit to the value of capital invested, regardless of the amount of labour it employs, or indeed whether or not that labour produces commodities at all. Marx offers a systematic account of the way in which competition among capitals gives rise to this linkage of profit with total capital invested by redistributing the surplus value created by productive labour.

Prices and the Collective Character of Exploitation (III, 9)

Because prices are not necessarily proportional to values, surplus value is not necessarily realized by the capitalists who hired the labour-power that created it. Exploitation is thus a 'collective'

mechanism for the capitalist class. It is as if surplus labour was collected in a single pool, and then distributed among capitalists in proportion to their invested capital (though the division of the surplus value among the individual capitals is actually the result of a fierce competitive struggle):

> Thus, although in selling their commodities the capitalists of the various spheres of production recover the value of the capital consumed in their production, they do not secure the surplus-value, and consequently the profit, created in their own sphere by the production of these commodities. What they secure is only as much surplus-value, and hence profit, as falls, when uniformly distributed, to the share of every aliquot part of the total social capital from the total social surplus-value, or profit, produced in a given time by the social capital in all spheres of production. . . . So far as profits are concerned, the various capitalists are just so many [100] stockholders in a stock company in which the shares of profit are uniformly divided per 100, so that profits differ in the case of the individual capitalists only in accordance with the amount of capital invested by each in the aggregate enterprise, i.e., according to his investment in social production as a whole, according to the number of his shares. (III, 9)

It is, consequently, necessary to distinguish between the mechanisms which govern the overall *appropriation* of surplus-value and its *realization* by particular capitalists:

1. The total surplus value depends on the value of labour power and the total number of workers capitalists employ.
2. Any system of commodity prices 'distributes' this total surplus value to individual producers (and landowners).

Marx describes this process of redistribution of surplus value as a 'metabolism' of value. Note that prices remain 'forms of value', as stated in the analysis of money and prices, but the hours of social abstract labour are reshuffled. At issue is no longer the labour actually expended to produce each commodity individually, but value as socially 'distributed' by prices (purchasing power as a fraction of social value 'conveyed' by the price of each commodity).

The Transformation Problem (III, 9)

At the beginning of Volume III, Marx pursues two objectives simultaneously. On the one hand, he analyses the basic mechanisms of competition in capitalism, in which the determination of a particular set of prices is implied, with equalized profit rates among industries, and, on the other hand, he uses this particular case to discuss the metabolism of value introduced above. This exposition obscures the fact that the underlying mechanism of exploitation operates whatever the prevailing system of prices; the theory of exploitation does not depend on the particular properties of commodity prices and, in particular, not on the attainment of a market equilibrium at which profit rates are equalized. The failure to separate the two projects, and to appreciate the restricted context of the discussion of the metabolism of value in this particular case, has created much confusion in the history of Marxist economic theory.

In the later literature the two problems, those of the metabolism of value and the prevalence of a particular set of prices in capitalist competition, are usually treated jointly as the *transformation problem.* Because of its importance in the history of Marxism, a specific entry is devoted to this controversial issue (see Marxian transformation problem).

The Classical Marxian Long-Period Equilibrium Prices of Production (III, 10)

The analysis of this process of redistribution of surplus value through competition marks an important break in the present account of Marx's analysis in *Capital.* Beginning with the definition of capital (and the corresponding requirement of the analysis of commodity and money, actually a preliminary to the exposition of capital), we first followed Marx in his investigation of the two components of the theory of capital: the extraction of surplus value and the circuit of capital. These two aspects were then combined in analyses such as the reproduction schemes or capitalist functions. Finally, attention turned to the division of surplus value: (1) its distribution as interest and dividends to money capitalists, and as rents to

landowners; (2) its realization by various categories of capitals, such as commercial capital and banking capital, in which no surplus value is produced; and (3), in the present section, its reallocation to capitalists of various industries independently of the extraction by individual capitalists, as in competition. We now enter a new category of developments, in which dynamic processes are involved: the mechanisms of competition, accumulation and employment, technical and distributional changes, and crises and the business cycle.

The basic idea in the analysis of capitalist competition is straightforward. If capital is free to move from one line of production to another in search of profit, the competitive movement of capitals will tend to move prices of specific commodities up or down until the rate of profit is equalized in all sectors. The *equalization of the rate of profit,* clearly stated by Adam Smith and David Ricardo, represents competition at the most fundamental level of analysis. The appropriation and realization of surplus value, as stated above, is thus specified quantitatively: one industry where little labour is used proportionally to total capital, in comparison to another industry, realizes more surplus value as profit than its workers actually contribute to the total surplus value (and conversely).

The *profit rate* is central in this analysis of competition. The profit rate is defined as the ratio of profit, s, to total capital, $K = M + C + P$, that is $r = s/K$. The ratio of the value of the average total capital invested during one unit of time (for example, a year) to the flow of value corresponding to the cost of production engaged during this unit of time, $T = K/(c + v)$, is the *turnover time* of capital measured in units of time such as months or years. In the Marxist literature, the turnover time is often implicitly or explicitly assumed to be unity, in which case the profit rate $r = s/K$ is equal to the *profit margin*, the ratio of profit to costs of production, $s/(c + v)$.

The movement of capital in the pursuit of profit results in a tendency toward the equalization of profit rates among industries. Marx calls commodity prices which are consistent with an equalized profit rate *prices of production:*

> But capital withdraws from a sphere with a low rate of profit and invades others, which yield a higher profit. Through this incessant outflow and influx, or, briefly, through its distribution among the various spheres, which depends on how the rate of profit falls here and rises there, it creates such a ratio of supply to demand that the average profit in the various spheres of production becomes the same, and values are, therefore, converted into prices of production. (III, 10)

Actual *market prices* tend to gravitate around prices of production, and this property defines the *capitalist law of exchange* (which supersedes the commodity law of exchange when production is organized by capital). As stated earlier, Marx calls the substitution of one law of exchange for the other a 'transformation', the *transformation of values* (actually prices proportional to individual values) *into prices of production.*

The Profit of Commercial and Money-Dealing Capital (III, 16; III, 19)

Although commercial and money-dealing capitals do not contribute to the extraction of surplus value, they do participate in its realization, along the lines indicated above, like any other capital. Commercial capital, for example, must secure a profit by buying commodities from industrial capitalists at prices below the prices at which those commodities will be sold to final purchasers. In this way commercial capital appropriates part of the surplus value actually created in the circuit of industrial capital. Similarly, the fees charged by money-dealing capital transfer surplus value created in the circuit of other capitals (abstracting from interest paid by other agents such as households or the state). Thus, the profit of commercial and money-dealing capital is part of the surplus value produced by labour employed by industrial capital.

Differential and Absolute Rent (III, 38; III, 45)

The level at which rents can be established is directly related to the level of the average and tendentially uniform profit rate in the overall economy. The condition for the cultivation of a land of lesser fertility or for a more intensive investment is that the marginal investment must yield the average profit rate. All capitalists

(including capitalist farmers) expect to realize the average profit rate prevailing throughout the economy. This condition is assured if landowners bargain for rents just high enough to assure capitalists the average rate of profit on their land. This defines *differential rent.* Marx also assumes that landowners as a class may withhold their lands until a minimum rent is paid, which defines *absolute rent.*

The Centralization and Concentration of Capital, Monopoly (I, 25)

The Classical–Marxian analysis, which assumes equalized profit rates among industries (not firms, because of differences in their productive efficiency), does not seem to match the features of competition in modern capitalism. Followers of Marx, from Hilferding and Lenin in the early 20th century to contemporary Marxist economics, point to the historical transformation of competition through the emergence of monopolies and oligopolies. The notion of the interplay of large firms is already part of Marx's analysis. In the process of accumulation the size of individual capitalist firms is altered by the *concentration* and *centralization* of capital. In Marx's account, concentration refers to the rise of the size of firms which parallels accumulation, while centralization denotes the outcome of merger or acquisition (and the process of competitive elimination of smaller and less efficient firms in an industry). Monopoly capital is not, however, part of Marx's analysis of capitalism, and Marx does not question the classical analysis of competition on such grounds. Rather than the view that the size of firms could hamper the process of equalization of profit rates among industries, Marx repeatedly asserts that credit mechanisms, including banks, are a crucial factor in the ability of capital to migrate among industries and, therefore, in the formation of prices of production.

Accumulation, and Technological and Distributional Change

The *accumulation of capital* refers to the situation where a fraction of surplus value is saved and devoted to increasing the value of capital. While the analysis of expanded reproduction considers a steady growth path of the economy (on which the key ratios, the rate of surplus value, the organic composition of capital, the value of labour power, and the composition of demand, are assumed to remain constant), Marx's theory of accumulation incorporates the qualitative change in capitalist production that actually accompanies its expansion.

Capital Accumulation and Employment (I, 25)

For accumulation to succeed, a number of conditions must be met. In particular, an expanded supply of labour power must be made available to permit the expansion of production, an issue which Marx addresses at the end of Volume I. Marx rejects the conclusions of classical economists such as Thomas Malthus, who proposed universal laws governing population growth and a 'natural' path of accumulation of capital, and blamed low wages on the fecundity of workers and the limits of natural resources. Marx argues that each mode of production evolves its own characteristic laws of population, and that capitalism in particular gives rise to a number of mechanisms that ensure a rough proportionality between population growth and the accumulation of capital.

How much labour is necessary to meet the demands of capital accumulation? How is the supply of labour roughly adapted to accumulation? Marx explains, in his *law of capitalist accumulation,* that the amount of labour required depends on (1) the pace of accumulation and (2) technical change as manifested in the variation of the composition of capital – that is, the ratio of capital outlays on means of production (*constant capital*) to capital outlays on wages (*variable capital*). If accumulation is rapid, and the composition of capital unaltered, the demand for labour power grows in proportion to accumulation and real wages tend to increase. This is the most favourable situation for workers. Technical change may moderate this tendency through an increase in the composition of capital, as the same accumulation requires less additional labour, and the demand for labour power grows

more slowly than capital as a whole. A priori, any relation between the pace of accumulation and the change in the composition of capital may occur. Marx points, however, to the fact that the composition of capital tends historically to rise and, thus, the pressure on employment is regularly relaxed.

Two mechanisms contribute to remedy any potential lack of available labour power. First, technical change leading to increases in the composition of capital makes some employed labour redundant. Second, recurrent crises periodically restore what Marx calls the *floating reserve army of labour,* with the decline of output.

Thus, the process of accumulation is uneven. Accumulation first proceeds during phases of more or less balanced growth; gradually the reserve army of unemployed workers is reabsorbed and wages rises. This is an inducement towards technical change increasing the composition of capital. If, however, the demand for labour grows too rapidly, a crisis occurs, the demand for labour is relaxed. Finally, a new wave of accumulation resumes after the crisis, during which a fraction of capital is devalued or destroyed. We will return below to these episodes in which a rise of wages provokes crises, which Marx calls situations of 'over-accumulation'.

In addition to this recurring fluctuation of unemployment, capitalism historically has drawn workers from the *latent reserve army*, through the destruction of traditional agricultural modes of production, and the consequent migration of displaced workers to the capitalist labour market. The potential competition of the latent reserve army puts a long-term downward pressure on wages as well.

The overall interaction of these factors is complex, because technical change and the income distribution cannot be treated as independent mechanisms. Marx considers that rising wages, and a correspondingly diminished rate of surplus value, increase the incentives for capitalists to seek labour-saving technical changes. This leads to a rise in the composition of capital, as more machinery is employed, precisely in order to avoid increased wage costs. This analysis must be supplemented by the consideration of fundamental political conditions, in particular, the strength of workers' class struggle, since Marx believed that, over and above the mechanisms involved in the law of capitalist accumulation, organized labour struggles could influence both wages and the length of the working day.

One of Marx's main goals in presenting his theory of accumulation, at the end of Volume I of *Capital,* is to show that the scarcity of labour power is not an absolute barrier to capital accumulation. The main thesis there is that, in the race between capital accumulation and the supply of labour power that governs the evolution of real wages, employment, and the rate of surplus value, capital has the edge over labour as a result of the capability of capital to substitute fixed capital (machinery) for labour:

> The same causes which develop the expansive power of capital, develop also the labour-power at its disposal. The relative mass of the industrial reserve army increases therefore with the potential energy of wealth. But the greater this reserve army in proportion to the active labour-army, the greater is the mass of a consolidated surplus-population, whose misery is in inverse ratio to its torment of labour. The more extensive, finally, the Lazarus-layers of the working-class, and the industrial reserve army, the greater is official pauperism. *This is the absolute general law of capitalist accumulation.* Like all other laws it is modified in its working by many circumstances, the analysis of which does not concern us here. (I, 25–4)

Besides the resistance of organized workers, this capability of capitalism to perpetuate an available reserve army by technical change is limited by the cost of the addition of capital which is required to displace labour, as Marx will contend in his analysis of technical change and the tendency for the profit rate to fall.

Technical Change (III, 13–15)

The social and technical conditions of production and their historical transformation are central to Marx's analysis of capitalist production. The term 'technology' is convenient but somewhat misleading. Marx always describes conditions of production in a perspective which combines technology in the strict sense and organization, that is, the institutional framework in which production is performed; the notion of social relations cannot be

neglected in this context. This is the case, for example, in the analysis of relative surplus value, as discussed earlier in reference to manufacture and large-scale industry.

Although Marx often discussed specific historical determinants of technical innovations, his main theory of technical change in capitalism sees it as an endogenous response to pressures from competitors and workers. Each capitalist has a strong motivation to find cost-reducing technical innovations (or profit-increasing product innovations) because the firm which first successfully exploits such innovations is in a position to capture higher-than-average profit rates ('super-profits') as a result of its temporary monopoly on the innovation. Innovating capitalists may also use their cost advantage aggressively to increase their market share. (In this respect Marx develops the theory of technical change Ricardo 1817, presents in his chapter on machinery.) Over time, competitors will find equivalent innovations and the advantage of the innovating capitalist will erode.

Capitalist technical innovation in Marx's framework begins with the discovery of a range of potential new productive techniques and forms of labour organization. The accumulated store of technical knowledge available to capitalist society at any moment is the result of this historical process of innovation: there is no set of predetermined techniques as is assumed in the neoclassical production function. Marx's theory of induced technical change is basically evolutionary. The capitalist evaluates the cost of these alternatives at prevailing prices and wages, and forms expectations concerning profit rates. Only those technologies that promise to reduce costs or increase profits at prevailing prices and wages are *viable* candidates for adoption. The criterion is an increased profit rate.

Marx emphasizes that, because capitalism places both strong incentives for technical change and the power to implement in the hands of competing capitalist firms, it is a *technically progressive* mode of production, in contrast to slavery and feudalism. In this respect Marx resembles Smith, who emphasizes increasing returns inherent in the division of labour, rather than Ricardo, who emphasizes diminishing returns due to limited natural resources (land).

The Tendency for the Profit Rate to Fall (III, 13–15)

In Volume III of *Capital,* Marx describes trajectories of technical and distributional changes that he denotes as *historical tendencies*. They are unbalanced (nonhomothetic) growth trajectories, which Marx considered typical of the dynamics of capitalism, which we will describe as *trajectories à la Marx.* Along such very long-term paths, the growth rates of capital, output, and employment gradually fall, labour productivity and the composition of capital rise, the share of wages in total income is constant or diminishing, and the profit rate declines. In the speaking of historical tendencies, 'historical' refers to a very long-term time frame; 'tendency' means that though accumulation in capitalism tends to follow such trajectories, the trajectory does not necessarily prevail due to the action of what Marx labels *counteracting factors.* It is in this framework that Marx defines the *tendential fall in the rate of profit.* This 'law' expresses sophisticated insights into the historical dynamics of capitalist economic growth. It is one of the major disputed issues in contemporary Marxist economics (along with the transformation problem).

In Volume III, the profit rate is written as a ratio of two flows or, equivalently, the turnover time of capital is assumed to be unity: $r = s/K = s/(c + v)$. Dividing by v, Marx obtains: $r = (s/v)/(c/v + 1)$. The numerator is the rate of surplus value, and the denominator is the *value composition of capital, the ratio of constant to variable capital,* plus 1. Marx calls this value composition the *organic composition* of capital. In this simple presentation, the conflicting impacts of the rate of exploitation and the organic composition of capital are clearly evident.

Although *labour productivity* does not appear in this formal setting, it is explicitly a key variable in Marx's analysis. Without altering the basic framework, it is possible to write: $r = (s/(v + s))/((c + v)/(v + s))$. Here, $s/(v + s)$ is the share of profit in total income, and $(c + v)/(v + s)$ is total capital per hour worked, which is another measure of the organic

M

composition of capital. (This ratio can also be read as the ratio of capital to output, since output is equal to total income, or equivalently the inverse of what is frequently loosely called 'capital productivity'.) The numerator, the share of profit, can be written $1 - (v/(v + s))$, that is, 1 minus the share of wages. The share of wages is equal to real wages divided by labour productivity. Thus, the profit rate can be expressed as the ratio of the profit share to the total capital per hour worked, which we call simply the *composition of capital*:

$$\text{profit rate} = \frac{1 - \dfrac{\text{real wage}}{\text{labor productivity}}}{\text{composition of capital}}$$

Marx's fundamental insight can be sketched as follows. To maintain or increase profits (which appear in the numerator of the profit rate), when there is no fall in the real wage, capitalists must increase the productivity of labour, which is the mechanism of relative surplus value. Marx contends, however, that this increase has a considerable cost for capitalists because increases in labour productivity typically require the investment of more capital per hour worked: productivity gains are realized by way of an increased mechanization of production. Thus, the composition of capital rises, and the rate of profit may fall. The actual evolution of the rate of profit also depends on what happens to the real wage and, consequently, to the rate of surplus value as labour productivity increases, which depends on labour market factors and class struggle, which are beyond the control of any individual capitalist.

Marx considers the case where the rate of surplus value remains constant to refute the argument that the falling profit rate is the result of an excessive growth in the cost of labour to the capitalists. When the productivity of labour rises, a constant rate of surplus value implies a rising real wage. Thus in making this argument, Marx does not assume a constant real wage. His thesis is rather that it is difficult for capitalists to counteract rising wages by technical change, since a more efficient technique in terms of labour productivity typically requires a rising composition of capital. The linchpin of Marx's thesis is, therefore, a hypothesis on the features of available techniques, that is, the profile of innovation: it is comparatively easy to find labour-saving devices if the cost of mechanization is not considered, but opportunities to reduce labour costs without inflating capital costs are rare.

Thus, on trajectories à la Marx the productivity of labour rises, while the productivity of capital (the inverse of the composition of capital) falls, a pattern of technical change sometimes called *Marx-biased*:

> The law of the falling rate of profit, which expresses the same, or even a higher, rate of surplus-value, states, in other words, that any quantity of the average social capital, say, a capital of 100, comprises an ever larger portion of means of production, and an ever smaller portion of living labour. Therefore, since the aggregate mass of living labour operating the means of production decreases in relation to the value of these means of production, it follows that the unpaid labour and the portion of value in which it is expressed must decline as compared to the value of the advanced total capital. ... The relative decrease of the variable and increase of the constant capital, however much both parts may grow in absolute magnitude, is, as we have said, but another expression for greater productivity of labour. (III, 13)

Though Marx never articulated the entire framework, this analysis of the biased pattern of technical change supplements the mechanisms at work in the law of capitalist accumulation. Accumulation recurrently pushes employment to the limits of the supply of labour power available and drives real wages upward. Technical change and recurrent crises allow for the partial relaxation of this pressure (as we have seen), but, in typical periods, the new techniques available are such that technical change can only partially offset the rise in real wages, and the profit rate falls. Accumulation is pursued in spite of the diminished profit rate, which will only be apparent after the fact, when a major crisis occurs.

The analysis Engels published from Marx's notes in Volume III of *Capital* is incomplete, and was not intended for publication in the form in which we read it. Consequently, it is not too surprising that Marx's analysis of the tendency for the profit rate to fall remains controversial among Marxists. A central issue is the assumption made concerning real wages, and its relationship to the

profitability criterion in the adoption of new techniques. Marx is clear that the innovating capitalist initially makes a surplus profit, while his competitors gradually adopt the new technique and prices fall through competition towards the prices of production corresponding to the new technology. Marx contends that the new uniform average profit rate tends to be lower than the original one. Nobuo Okishio (1972) has demonstrated that if the real wage remains unchanged during this process the new average profit rate can never fall. But along a trajectory à la Marx real wages do increase, as we have explained, although the possibility of a tendency for the rate of profit to fall is consistent with Marx's assumption that the rate of surplus value is constant or even rising.

The problem of the evolution of real wages, the value of labour power, and the rate of surplus value over time as labour productivity rises is controversial among Marxists, due to a change of Marx's view on this subject during his lifetime. Engels explained that Marx originally accepted the so-called *iron law of wages,* which assumes that real wages are constantly driven downward to a minimum compatible with the reproduction of the labour force, but later abandoned it. Marx sometimes refers to a 'socially and historically determined' cost of reproduction of labour power, as an external constraint on the evolution of the real wage. But this 'exogenous' variable is explicitly subject to a number of economic and social determinations: (i) class struggle impacts on wages and the duration of labour; and (ii) the outcome of struggles crucially depends on the conditions of accumulation and the population available to work (as in the law of accumulation). Marx's understanding of the determination of wages is similar to his view of technical change: the path of real wages is the result of the interaction of extra-economic factors with economic mechanisms such as accumulation and crises.

Crises and the Business Cycle (III, 15)

There is no systematic treatment of crises and of the business cycle in Marx's work, although the issue plays a prominent role in his analysis of capitalism. In early works, like the *Communist Manifesto,* even prior to Marx's serious study of political economy, the idea that crises will prove more violent with the evolution of capitalism is central. Recurrent crises became a feature of capitalism during the first half of the 19th century. This link between economic mechanisms and class struggle had a considerable impact on Marx's view of the historical dynamics of capitalism. Then, Marx became gradually better aware of the complexity of the phenomenon of crises, in particular the relationship between real and financial mechanisms and crises.

Partial Crises and Crises of General Overproduction

Before capitalism, poor crops and the devastation of war and disease were the major causes of disruptions of production. David Ricardo (1817) observed the existence of recurring crises more directly related to the nature of capitalism, which he called *states of distress.* These crises struck specific industries, like textiles. Consequently, Ricardo interpreted these situations as the effect of *disproportions*, that is, the outcome of the excessive accumulation of capital in one industry. Ricardo did not believe in the possibility of a general glut of the market. Marx devoted much energy to the refutation of Ricardo's interpretation. He contended that the existence of a delay between the sale of a commodity and the spending of its money price on another commodity invalidates 'Say's Law', the principle that the sale of a commodity constitutes a direct demand for another commodity. Monetary exchange thus implies the *possibility* of crises, because, by functioning as intermediary in exchanges, money allows for the interruption of the chain of exchanges. Only the 'possibility' of crises is, however, implied, not their actual mechanisms in capitalism.

Marx identified a new category of crises, *crises of general overproduction,* where all industries were simultaneously affected. Marx did not deny the existence of crises specific to particular industries, that he called *partial crises*, but contrasted the two types of situations, partial and general, and was specifically concerned with the latter.

The Ultimate Ground of Crisis. Profitability and Social Needs

Marx described general crises of overproduction as typical of capitalism. In capitalism, the purpose of production is not the satisfaction of the needs of the population, but the appropriation of profits. The 'ultimate ground' of crisis in capitalism is this disconnection between production and social needs:

> The ultimate reason [*ground*] for all real crises always remains the poverty and restricted consumption of the masses as opposed to the drive of capitalist production to develop the productive forces as though only the absolute consuming power of society constituted their limit. (III, 30)

This quotation is often misunderstood. Marx did not believe that higher wages would solve the problem of crises in capitalism. The cause of crises, proper to capitalism, is the recurrent inability to pursue production *at a certain rate of profit*. Therefore, profitability is always the crucial variable in Marx's explanation of crises:

> Over-production of capital is never anything more than overproduction of means of production – of means of labour and necessities of life – which may serve as capital, *i.e.,* may serve to exploit labour at a given degree of exploitation; . . . too many means of labour and necessities of life are produced at times to permit of their serving as means for the exploitation of labourers at a certain rate of profit. (III, 15–3)

The Business Cycle and Its Determinants

Marx described the fluctuating pattern of production in capitalism as 'the cycles in which modern industry moves – state of inactivity, mounting revival, prosperity, over-production, crisis, stagnation, state of inactivity, etc.' (III, 22).

Production is recurrently destabilized by mechanisms which affect the profitability of capital in the short run (a sudden decline rather than a steady downward trend). The first mechanism is *over-accumulation*. Periodically, employment gets closer to the limits of the population available to work (the reserve army is reabsorbed, as in the law of capitalist accumulation). Wages tend to rise, and profitability is diminished. A second mechanism is the rise of interest rates. During the phase of rapid accumulation, the mass of credits increases and, at a certain point, interest rates rise. Again, profitability is affected and the economy destabilized. Marx is well aware of the relationship between real and financial mechanism, and he interprets the direction of causation as reciprocal.

As stated above, Marx did not explain crises by the deficient level of wages (except in his very early work), and refuted this explanation in the manuscripts of Volume II:

> It is sheer tautology to say that crises are caused by the scarcity of effective consumption, or of effective consumers. The capitalist system does not know any other modes of consumption than effective ones, except that of *sub forma pauperis* or of the swindler. That commodities are unsalable means only that no effective purchasers have been found for them, i.e., consumers (since commodities are bought in the final analysis for productive or individual consumption). But if one were to attempt to give this tautology the semblance of a profounder justification by saying that the working-class receives too small a portion of its own product and the evil would be remedied as soon as it receives a larger share of it and its wages increase in consequence, one could only remark that crises are always prepared by precisely a period in which wages rise generally [over-accumulation] and the working-class actually gets a larger share of that part of the annual product which is intended for consumption. From the point of view of these advocates of sound and 'simple' (!) common sense, such a period should rather remove the crisis. (II, 20)

Structural Crises and the Falling Profit Rate

Since the profitability of capital is central in Marx analysis of crises, there is a link between the tendency for the profit rate to fall and crises. Marx's view is that actual phases of decline of the profit rate make crises more likely, more frequent and deeper. He points to the existence of periods of sustained instability, which, although Marx does not use the term, can be called *structural crises.* A declining and depressed profit rate (both the tendency and levels are at issue) disturbs capitalist accumulation:

> . . . in view of the fact that the rate at which the total capital is valorised, i.e. the rate of profit, is the spur to capitalist production . . ., a fall in this rate slows down the formation of new, independent capitals and thus appears as a threat to the development of the capitalist production process; it promotes overproduction, speculation and crises, and leads to the existence of excess capital alongside a surplus population. (III, 15)

This insight concerning the link between the profit rate and the occurrence of periods of

historical perturbation in the course of accumulation provides a powerful framework for understanding the real history of capitalist economies.

See Also

- British Classical Economics
- Capitalism
- Class
- Classical Distribution Theories
- Classical Growth Model
- Commodity Fetishism
- Commodity Money
- Exploitation
- Labour Theory of Value
- Labour's Share of Income
- Marx, Karl Heinrich (1818–1883)
- Marxian Transformation Problem
- Marxian Value Analysis
- 'Political Economy'
- Profit and Profit Theory

Bibliography

Arthur, C.J. 2004. *New dialectic and Marx's capital*. Leiden/Boston: Brill.

Brewer, A. 1984. *Guide to Marx's capital*. Cambridge: Cambridge University Press.

Cleaver, H. 2000. *Reading capital politically*. Edinburgh: AK Press.

Duménil, G., and D. Lévy. 1993. *Economics of the profit rate: Competition, crises, and historical tendencies in capitalism*. Aldershot: Edward Elgar.

Duménil, G., and D. Lévy. 2003. *Économie Marxiste du Capitalisme*. Paris: La Découverte.

Fine, B. 1975. *Marx's capital*. Leiden/Boston: Brill.

Fine, B., and A. Saad-Filho. 2004. *Marx's capital*. New York: Pluto.

Foley, D. 1986. *Understanding capital: Marx's economic theory*. Cambridge, MA: Harvard University Press.

Itoh, M., and P. Bullock. 1987. *The basic theory of capitalism: The forms and substance of the capitalist economy*. Lanham: Rowman and Littlefield.

Marx, K. 1859. *A contribution to the critique of political economy,* ed. M. Dobb. New York: International Publishers, 1970.

Marx, K. 1976, 1978, 1981. *Capital, Volumes I, II, III*. New York: Random House.

Marx, K., and F. Engels. 1848. *The communist manifesto*. New York: Signet Classics, 1998.

Moseley, F. 1993. *Marx's method in capital: A reexamination*. Leiden/Boston: Brill.

Okishio, N. 1972. On Marx's production prices. *Keizaigaku Kenkyu* 19: 38–63.

Pilling, G. 1980. *Marx's capital*. London/New York: Routledge.

Ricardo, D. 1817. *On the principles of political economy and taxation*, 1951. Cambridge: Cambridge University Press.

Rosdolsky, R. 2005. *The making of Marx's capital*. London: Pluto Press.

Rubin, I. 1972. *Essays on Marx's economic theory*. Detroit: Black and Red.

Sekine, T. 1997. *Outline of the dialectic of capital*. London: Palgrave Macmillan.

Shaikh, A., and E. Tonak. 1996. *Measuring the wealth of nations: The political economy of national accounts*. Cambridge: Cambridge University Press.

Sweezy, P. 1970. *Theory of capitalist development*. New York: Monthly Review Press.

Weeks, J. 1994. *Capital and exploitation*. Princeton: Princeton University Press.

Wolff, R., and M. Cohen. 1985. *Understanding Marx: A reconstructive and critique of capital*. Princeton: Princeton University Press.

Marxian Transformation Problem

Duncan Foley and Gérard Duménil

M

Abstract

The transformation problem relates the labour theory of value and the competitive equalization of the rate of profit. Marx distinguishes the production of surplus-value from its redistribution through prices. Critics claim that the labour theory of value is an unnecessary detour to the determination of prices because total value and surplus-value are not conserved. The Single-System Labour Theory of Value (SS-LTV) argues that at any prices (1) the price of the net product expresses the labour expended, and (2) total profits are the price form of surplus-value, because the value of labour-power is the labour time equivalent of the wage.

Keywords

Capitalist law of exchange; Commodity law of exchange; Constant and variable capital; Dual system; Exploitation; Fundamental Marxian

Theorem; Gravitation; Labour power; Labour theory of value; Labour time; Marx, K. H.; Marxian transformation problem; Monetary expression of value or labour time; Natural price; Net product; 'New interpretation' of labour theory of value; Profit, rate of; Ricardo, D.; Single-system labour theory of value; Smith, A.; Surplus-value; Transformation problem; Value; Value added; Value of labour-power; Variable capital

JEL Classifications
B1

Marx's Framework: Value, Surplus-Value, Prices and Competition

Marx consistently distinguishes the notions of *value* and *price*, in contrast to contemporary economic language, which uses the term 'value' to refer to prices in a situation of general equilibrium, though the use of the term is rather flexible; for example 'value added' is actually the value of net product measured in price terms. For Marx, value is a 'social substance' manifested in economic relations in the 'form' of prices, though prices are not necessarily proportional to values, as we will see.

Value and Surplus-Value

We first recall Marx's basic concepts (see also Marx's analysis of capitalist production). Central to Marx's framework of analysis in *Capital* is the *labour theory of value* (LTV), which defines the value of a commodity as the 'socially necessary' labour time required by its production, that is, the labour time required by average available techniques of production for workers of average skill.

The LTV is central to Marx's theory of exploitation, a term he uses to describe a situation in which one individual or group lives on the product of the labour of others. According to the LTV, when commodities are exchanged through sale and purchase, no value is created. But this principle does not apply to capitalists' purchase of the *labour power* of workers. Workers sell their labour power, that is, their capability to work, to a firm, owned by a capitalist. The buyer uses this labour power in production to add value to the commodity produced. The *value of labour power* is the labour time required by the production of the commodities the worker buys. But the worker can typically work more hours than are on average required to produce this bundle of commodities. For example, the goods the worker can buy may require eight hours of labour per day, when the labour-day lasts 12 hours. The difference, four hours, is unpaid labour time. If an hour of social labour on average produces a value whose price form is \$10, four hours of unpaid labour time results in a surplus-value whose price form is \$40, which is appropriated by the capitalist. The *rate of surplus-value* is the ratio of unpaid to paid labour time, in this case 4/8, that is, 50 per cent.

Two Laws of Exchange

Marx situates his discussion in the context of the distinction made by Adam Smith and David Ricardo between 'market prices' and 'natural prices'. Market prices are the prices at which commodities actually exchange from day to day in the market. Smith and Ricardo, however, regarded market prices as fluctuating (or 'gravitating') around centres of attraction they called 'natural prices'. ('Gravitation' means that the economy is in a permanent situation of disequilibrium, though in a vicinity of equilibrium where natural prices would prevail.)

In the above analysis, Marx assumes that commodities tend to exchange at their values (at *prices proportional to values*), that is, in proportion to the labour time embodied in them. 'Tend' means here that deviations are obviously possible, but that such prices will 'regulate' the market, in the sense that if the prevailing set of prices systematically under-compensates the labour used in the production of a commodity, labour will move to the production of better-paid commodities. As a result, the supply of the under-compensated commodity will decline, and its price will rise. In reality prices would gravitate around values, which would play the role of natural prices in such an economy. This is the *commodity law of exchange*.

In a capitalist economy, however, capitalists buy not only the labour power of workers (which Marx denotes as *variable capital*), but also non-labour inputs, such as raw materials, and fixed capital, such as machinery (which Marx denotes as *constant capital*). If natural prices were proportional to labour inputs, as the commodity law of exchange posits, capitalists using more constant capital per worker than the average would realize smaller profit in comparison to their total capital advanced, that is, lower profit rates. Marx accepts the idea that competition tends to equalize profit rates in various industries, despite differences in capital advanced per worker, which is the *capitalist law of exchange*. Marx uses the term 'prices of production' to describe a system of prices which guarantee to the capitalists of various industries a uniform profit rate. Capitalists will invest more where profit rates are larger, and conversely in the symmetrical case. They move their capital from one industry to another seeking maximum profit rates, and this movement results in a gravitation of market prices around prices of production. Marx regards prices of production as the centres of gravitation of market prices, and thus the natural prices relevant to a competitive capitalist economy.

Is the Theory of Surplus-Value Compatible with the Theory of Competition?

The problem is posed of the compatibility of the capitalist law of exchange at prices of production with the theory of exploitation as extraction of surplus-value. Marx's line of argument is that surplus-value is *created* in production through the exploitation of labour, that is, in proportion to labour expended, but *realized* proportionally to total capital invested. According to Marx, this separation between the locus of extraction and the locus of realization does not contradict the theory of exploitation so that capitalist competition is compatible with his theory of exploitation through the appropriation of surplus-value from unpaid labour time.

To support this argument, Marx presents a pair of tables (1981, ch. 9) showing the redistribution of surplus-value through deviations of price from values proportional to embodied labour times. All variables are measured in hours of labour time, and as a result prices of production are expressed in the same unit. Because Marx's own calculations involve some extraneous complexity (differential turnover rates among sectors), it is more useful to consider the simplified case shown in Table 1. Two industries exist, each of which advances the same capital of 100, but divided in different proportions between the purchase of non-labour inputs (C) and labour inputs (V). All capital is used up during the period, so that the rate of profit is the ratio of surplus-value to total capital advanced, $r = s/(c + v)$. The rate of surplus-value is uniform and equal to 100 per cent. Consequently, surplus-values are equal to variable capitals. Surplus-values and values are computed in each industry. When prices are proportional to values, profit rates differ between the two sectors. Prices of production are determined in Marx's procedure by summing up all surplus-value, a total of 40, and redistributing it in proportion to total capital, that is 20 in each industry, to equalize profit rates on the capitals advanced.

The procedure illustrates a straightforward 'redistribution' of surplus-value. Clearly, the sum of prices, 240, is equal to the sum of values, and total surplus-value is, by construction, conserved in the form of profit. These observations are expressed in two *Marxian equations* concerning the entire economy:

$$\textit{Sum of values} = \textit{sum of prices of production}$$
$$\textit{Sum of surplus value} = \textit{sum of profits}$$

Note that these compact formulations are not rigorous, since values and surplus-value are measured in labour time and prices and profits in money. Thus, 'Sum of values' should read 'Sum of prices proportional to values'. A simple way out of the problem of units is to use one of these equations to define the general level of prices. For example, the sum of prices of production could be set equal to the number of hours corresponding to the sum of values. Then, Marx's line of argument implies that the surpluses in both sets of prices are equal, as in the second equation. This simple

Marxian Transformation Problem, Table 1 Marx's calculation of prices of production from values

Industry	Constant capitals, C	Variable capitals, V	Total capitals, $K = C + V$	Surplus-values, $S = V$	Values of commodities produced, $\Lambda = K + S$	Profits, Π	‹Prices of production› of commodities produced, $P = K + \Pi$
1	70	30	100	30	130	20	120
2	90	10	100	10	110	20	120
Total economy	160	40	200	40	240	40	240

calculation illustrates the idea that profits are 'forms' of surplus-value, that is, unpaid labour.

Approximations

Marx is, however, aware that the type of computation illustrated in Table 1 is not satisfactory, since the evaluations of constant and variable capital have not been modified despite the fact that prices have changed.

First, when natural prices are prices of production, non-labour inputs are purchased on the market at prices of production, not at prices proportional to values. It is, therefore, not correct to conserve the evaluation of constant capital:

> We had originally assumed that the cost-price of a commodity equalled the value of the commodities consumed in its production. But for the buyer the price of production of a specific commodity is its cost-price, and may thus pass as cost-price into the prices of other commodities. Since the price of production may differ from the value of a commodity, it follows that the cost-price of a commodity containing this price of production of another commodity may also stand above or below that portion of its total value derived from the value of the means of production consumed by it. It is necessary to remember this modified significance of the cost-price, and to bear in mind that there is always the possibility of an error if the cost-price of a commodity in any particular sphere is identified with the value of the means of production consumed by it. Our present analysis does not necessitate a closer examination of this point. (Marx 1981, ch. 9)

Second, there is a similar problem concerning variable capital. When commodities exchange at prices of production, workers will not be able to buy the same bundle of commodities with a wage corresponding to a purchasing power expressed, as in Marx's calculation, as a certain number of hours of labour time, as when prices are proportional to values. Marx is also aware of this problem:

> [. . .] the average daily wage is indeed always equal to the value produced in the number of hours the labourer must work to produce the necessities of life. But this number of hours is in its turn obscured by the deviation of the prices of production of the necessities of life from their values. However, this always resolves itself to one commodity receiving too little of the surplus-value while another receives too much, so that the deviations from the values which are embodied in the prices of production compensate one another. Under capitalist production, the general law acts as the prevailing tendency only in a very complicated and approximate manner, as a never ascertainable average of ceaseless fluctuations. (Marx 1981, ch. 9)

It is not easy to understand Marx's position from these notes (which he never revised for publication). It does seem that the analysis requires a 'closer analysis', since the revaluation of constant capital at prices of production will in general make the sum of prices of production deviate from the sum of values, or make the sum of profits deviate from the sum of surplus-values. While it is true that a redistribution of surplus-value through a system of prices of production does not alter the living labour expended in production, so that over the whole economy the deviations from value 'compensate one another', the value of labour power will remain constant only if workers consume commodities in the same proportion as they are produced in the whole economy, which is implausible. The phrase 'average of ceaseless fluctuations' suggests the averaging out of market prices to prices of production rather than the averaging of surplus-value across sectors.

If Marx's use of the term 'approximately' is taken literally, it would appear that the LTV

and the theory of exploitation he introduced in volume 1 of *Capital* are only 'approximately' true! Although Marx is conscious of the problem, it is impossible to consider his solution as rigorous. In the formulation of the two equations above, it appears that, when the calculation is done rigorously as in the formal setting below, the second equation *does not hold*! Later critics have judged this a devastating refutation of Marx's theories of value and exploitation, which in turn has led to ongoing controversy.

Earlier Approaches

The foundations of the transformation problem can be found in the first analyses of competition and prices in capitalism, beginning with Adam Smith and David Ricardo, on which Marx elaborated. The distinction between values and prices remains somewhat fuzzy in these authors. Smith fails to establish a clear relationship between value and profit rate equalization as the principle determining 'natural prices'. Thus, one characteristic feature of these approaches, from which Marx was unable to depart completely, is that two sets of prices (the two laws of exchange above) are considered, one proportional to values (embodied labour times), and the other equalizing profit rates (a *dual system*), when only one price system prevails in real-world capitalism (a *single system*):

1. A system of prices proportional to values (embodied labour times) plays a role in the analyses of Smith, Ricardo and Marx. Only Marx, however, clearly distinguishes the two systems from the start.
2. The determination of the 'surplus', when such a concept exists (as in Ricardo and Marx), is posed in the first system and imported into the second, instead of being analysed directly within the second system.

This dual system approach lies at the basis of the phrase 'transformation problem', which refers to the transformation from one system into the other.

Adam Smith

Smith's point of departure is an 'early, rude' state of society, before the establishment of private property in land and means of production. There, Smith contends, products of human labour will exchange in proportion to the labour time required to produce them. Smith offers as an example that, if it requires two days on average to kill a beaver, but one day to kill a deer, a beaver will tend to exchange for two deer. Smith's argument supporting this conclusion rests on the assumption that any hunter can choose to allocate time to hunting deer or beaver, so that, if the exchange ratio were higher or lower than the labour time ratio, hunters would shift from the underto the over-remunerated productive activity, and force the exchange ratio back toward the labour time ratio. The viewpoint is clearly that of the commodity law of exchange.

Smith applies the same type of reasoning to argue that, once means of production have become private property (which he calls 'stock', and later economists called 'capital'), the ability of owners to shift their capital from one line of production to another will tend to equalize the profit rate across different sectors of production. The viewpoint is now that of the capitalist law of exchange.

David Ricardo

Ricardo critiques and corrects Smith's analysis. Ricardo originally based his theories of prices and distribution on Smith's first principle that the labour expended in producing a commodity determines its price in exchange. But Ricardo, elaborating on the dual system approach, examines the necessary quantitative difference between the two principles that might determine natural prices more carefully than Smith. Ricardo understood that the proportion between capitals invested in non-labour inputs and labour is not uniform across industries, and that this fact implies a discrepancy between the two sets of prices, but he regarded these deviations as quantitatively limited. Prefiguring Marx's investigation, Ricardo was concerned to work out the properties of the first system (values) to derive conclusions concerning

distribution, which he supposed were also valid in the second system (prices of production).

First, when natural prices are proportional to values (embodied labour times), it is obvious that there is a trade-off between the shares of output which respectively go to workers and capitalists: workers create all the value added to inputs, and buy a share of output whose production requires less labour time than they expend. In contrast to Smith, Ricardo had a clear view of this mechanism. This division of total output between workers and capitalists was crucial to his analysis, because of its implications in terms of economic policy. (For example, Ricardo was in favour of a low price of corn, which, in his opinion, would increase the profits of capitalists by lowering wages – and encourage capital accumulation.)

Second, Ricardo would have liked to conserve the straightforward distributional properties he derived from the assumption of prices proportional to values, even while acknowledging the quantitative difference between such natural prices proportional to values and natural prices that would equalize profit rates across industries. But Ricardo understood that, in the profit rate-equalizing system, the natural prices of commodities may change with a change in the real wage (due to the distinct compositions of capital) even if the labour required in production remains unaltered, contrary to what happens in the first system, where values remain unchanged with a change in the wage. Thus, with Ricardo's analysis, we are getting closer to Marx's framework and problems.

The Rebellious Classical Legacy in Marx

Marx adopted key elements from Smith and Ricardo's works: (*a*) a dual system approach to natural prices in capitalism (beginning, with Smith, as if labour was the unique input); (*b*) Ricardo's analysis of distribution as a 'trade-off' between wages and profit; and (*c*) Smith's analysis of competition that Ricardo had also adopted.

The two classical economists were the mainstream when Marx started his study of economics. Marx seized this opportunity to establish his theory of exploitation, in which surplus-value arises from unpaid labour time, on 'mainstream' grounds. Then he devoted hundreds of pages (in the manuscripts known as *The Theories of Surplus-value*) to the inability of these 'bourgeois' economists to establish a theory of exploitation, although Ricardo came close. This very smart political move on Marx's part eventually forced mainstream economic theory to abandon these 'dangerous' implications of the LTV.

The Transformation Controversy

A large literature is devoted to the transformation problem, starting with the critical contributions of Eugen von Böhm-Bawerk (1890) and Ladislaus von Bortkiewicz (1952) in the late 19th and early 20th centuries. This literature has led to considerable formal advance, though it has failed to resolve the basic controversy over which of Marx's conclusions, if any, are logically valid.

There are fundamentally two points raised by these critiques. First, the critics claim that the value system is useless as a preliminary to the calculation of prices of production. Paul Samuelson puts this point in the following manner: 'Contemplate two alternative and discordant systems. Write down one. Now transform by taking an eraser and rubbing it out. Then fill in the other one. *Voilá!* You have completed your transformation algorithm' (Samuelson 1971, p. 400). This point is, however, not really relevant, since Marx's objective was not to show that it is impossible to compute prices of production if values have not been previously determined, but rather to show that the theory of exploitation is consistent with the principle of capitalist competition.

Second, the main focus of this critique is the incompatibility of the two Marxian equations. This literature calculates surplus-value by deducting the value of a given bundle of worker's consumption from the worker's labour time. Profits, on the other hand, are calculated by deducting the price of this same bundle at prices of production from the value added (in prices). When prices of production are not proportional to values, these two quantities are not equal, violating the second Marxian equation. This treatment of the wage of workers, which allocates their

purchasing power to particular commodities, departs from Marx's apparent stipulation in his discussion of the transformation problem of the rate of surplus-value.

In face of this quantitative inequality between surplus-value and profit, the Fundamental Marxian Theorem (see Morishima 1973) argues that the LTV does provide a qualitative foundation for Marx's theory of exploitation, since the rate of profit will be positive if and only if the rate of surplus-value is positive. This interesting observation, however, falls short of fulfilling Marx's ambition to found his theory of exploitation on the LTV through the two Marxian equations.

A crucial moment in the criticism of Marx's transformation was the publication of Piero Sraffa (1960). This book is simultaneously a critique of Marx and of neoclassical economics, but it is, above all, a bold attempt to elaborate Ricardo's analysis. It is the origin of the neo-Ricardian school, represented by, in particular, Ian Steedman (1977) and Pierangelo Garegnani (1984). The central point, in the neo-Ricardian School, is that the LTV is useless, with respect to both the determination of prices of production and exploitation. The dual-system approach of Ricardo is abandoned in favour of the price of production system, as the reference to value is deemed irrelevant. Sraffa calculates prices of production directly from a description of technology and distribution. In this framework, he shows that Ricardo's trade-off between wages and the profit rate can be derived formally as a downward sloping relation (see the mathematical section below).

The Price of Net Product-Unallocated Purchasing Power Labour Theory of Value (PNP-UPP LTV) Approach to Exploitation

In the late 1970s, Gérard Duménil (1980, 1983, 1984) and Duncan Foley (1982) (independently) proposed new lines of interpretation of Marx's theory of value. In doing so, they followed distinct routes, but the basic principles underlying these reformulations converge to the same basic framework. This interpretation is inappropriately referred to, in the literature, as the 'New Interpretation'. It is more precise to describe it as the 'price of net product-unallocated purchasing power labour theory of value' (PNP-UPP LTV). It was rapidly adopted by Alain Lipietz (1982).

Value and Exploitation in the PNP-UPP LTV Approach

Beginning with Marx's two equations, as is traditional, there are two basic principles to this interpretation. First, Marx's equation concerning the 'sum of values' and 'sum of prices' holds for the net product of the period. 'Net product' means here, as in Marx's reproduction schemes and national accounting frameworks, output minus non-labour inputs inherited from the previous period. The important idea here is that it is the expenditure of living labour that creates value. Marx regards the value of a commodity as equal to the value transferred by the inputs consumed and the new value created by labour during the period. But the two perspectives are equivalent:

$$\begin{aligned} &\textit{Value transferred from inputs} \\ &+ \textit{value created by new labour} \\ &= \textit{value of output Value created by new labour} \\ &= \textit{value of output} \\ &- \textit{value transferred from inputs} \end{aligned}$$

The price form of the value created by the total productive labour expended during a period of time is the price of the net product of the period. (As is well known, the price of this net product is equal to total income, wages plus profits.) The PNP-UPP LTV interpretation argues that, when Marx (in the first quotation above) points to the fact that the cost-prices of commodities used as inputs to production must be adjusted to reflect the change to prices of production, the correct formulation would have been to exclude them from the first Marxian equation, which would then read 'Sum of values of net product = sum of price of net product'. Since values are expressed in labour time, while prices of production are expressed in terms of money, this equation implicitly defines an equivalence between value-creating labour time

and money, the *monetary expression of value or labour time* (MELT), which is the ratio of the price of net product (value added measured in money) to the productive labour time expended. If, for example, 250 billion hours of productive labour were expended in an economy to produce a net product worth $10 trillion, the monetary expression of labour time would be $40 per hour. The MELT expresses quantitatively (as a ratio of the price of the net product to the living labour expended) what Marx calls the 'price form' of the total value created during the period.

Second, the PNP-UPP LTV views the term 'surplus-value' in the second Marxian equation as referring to the monetary equivalent of unpaid labour time. The wage, as in Marx's calculation, is regarded as unallocated purchasing power giving workers the potential to buy a fraction of the net product. (This is the way capitalists look at wage payments, since the individual capitalist has no interest in how workers actually spend their wages.) Individual workers can allocate this purchasing power among the commodities they jointly produced (or even save some of it), in whatever proportions they choose. This can be described as the unallocated purchasing power (UPP) approach to exploitation. With this definition of surplus-value, the Marxian second equation immediately holds as an identity. The PNP-UPP LTV holds the rate of surplus-value rather than the consumption bundle of workers constant.

There is a sharp contrast between the PNP-UPP LTV and the traditional interpretation in the way they conceptualize distribution. Following Marx's procedure in his calculation, represented in the simplified example introduced earlier, it is impossible to assume that workers can buy the same bundle of commodities before and after the redistribution of surplus-value, since the purchasing power they receive will be spent at different prices. Consequently, the wage must be changed to keep the bundle of workers' consumption unchanged (and the rate of surplus-value must be altered – hence the controversy). The UPP approach to exploitation conserves the rate of exploitation, or, more rigorously, measures the value of labour power as the value whose price form is the price of the commodities workers can buy: an unallocated purchasing power on any commodities. The rate of surplus-value, as in Marx's calculation, is unchanged.

A Single-System Approach and Exploitation in any Set of Prices

A key aspect of the PNP-UPP LTV interpretation is that value is present in the theory of exploitation, as a social substance extracted in one place in the economy (firm, industry), and realized in another. But there is no logical anteriority in the value system, compared to the price system. This interpretation is a single-system approach to the LTV.

This property has important analytical consequences. There is only one economy, one system, not two. There is no 'underlying', hidden economy, which operates in 'values' where the distributional realities that structure the functioning of capitalism could be determined. The theory of exploitation is not dependent on the prevalence of any particular set of prices. The consideration of prices of production is not central to Marx's argument concerning exploitation, only an example that illustrates a much more general conclusion. Prices of production are just *one case* in which such a demonstration must be made, which Marx focused on because of the importance of this particular set of prices in competitive capitalism, as centres of gravitation of market prices.

The specific property expressed in the equality of the profit rate among industries cannot play any role in the theory of exploitation. Prices may deviate from prices of production because of gravitation; the amounts of surplus-value realized in each industry may also differ from what is implied by the prevalence of uniform profit rates because of the existence of non-reproducible resources and their rents; counteracting factors, such as monopoly, may also prevent equalization of profit rates. These deviations, inherent to capitalism, and also mentioned in Marx's analysis, do not invalidate his theory of value and exploitation.

An Ongoing Debate

The shift of perspective to single-system interpretations of Marx's labour theory of value has

led to further debate in this vein. Fred Moseley (2000) proposes to apply the reasoning of the SS-LTV approach not just to variable capital, but to constant capital as well. Moseley argues for retaining the original form of the Marxian equations by defining the total value of a commodity as the labour-time equivalent of the price of constant capital plus the living labour expended in adding value. Moseley argues that Marx's comments in the quotations above are unnecessary because Marx's tables themselves express his underlying understanding of the labour theory of value.

Alan Freeman, Giugelmo Carchedi, Andrew Kliman, and their co-authors (Freeman and Carchedi 1996) have put forward a 'temporal single-system' (TSS) interpretation of the labour theory of value. This interpretation sets the transformation problem in a temporal context, defining the value of commodities as the sum of the labour time equivalent of constant capital (calculated using a monetary expression of labour time) and the living labour expended in the current period in production. By construction, this interpretation makes the first Marxian equation hold for the total product, while the second Marxian equation holds when the monetary expression of labour time is appropriately defined (as in the SS-LTV). It is, however, clear in Marx's analysis that the value of a commodity is not determined by the actual amount of labour its production required in the past, but by the labour time it requires under present prevailing conditions:

> . . .the value of commodities is not determined by the labour-time originally expended in their production, but by the labour-time expended in their reproduction, and this decreases continually owing to the development of the social productivity of labour. On a higher level of social productivity, all available capital appears, for this reason, to be the result of a relatively short period of reproduction, instead of a long process of accumulation of capital. (Marx 1981, ch. 24)

This evaluation at 'replacement costs', however, does not imply that the economy is necessarily in a stationary state as the TSS critique has claimed.

A Mathematical Setting

The use of numerical examples to work out the quantitative implications of theoretical ideas is now outdated. The most common framework in the contemporary literature on the transformation problem is a pure circulating-capital model with a single technique in each sector, in which basic properties of solutions and interpretations can be elegantly and compactly expressed. A single homogeneous labour input works with stocks of an arbitrary but finite number of produced commodities available at the beginning of a production period. One unit of each commodity is produced by a single technique of production. This framework is consistent with the example in the first table above but not with Marx's tables since the circulating capital model does not include fixed capital, while Marx's examples do.

1. *Techniques of production*. The number of goods is n, also the number of techniques. A technique of production, indexed by j, is characterized by a column vector, $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{ji}, \ldots, a_{jn})$, and a scalar l_j, where a_{ji} is interpreted as the quantity of the commodity i required as inputs, and l_j as the quantity of labour required for the production of one unit of commodity j. A technology consisting of the set of all available techniques is described by collecting corresponding inputs into a matrix $\mathbf{A}$, and the labour input scalars into a row vector $\boldsymbol{l}'$. A pattern of economic production is described by a vector of levels of operation of the techniques, $\boldsymbol{x} = (x_1, 1\ldots, x_j, \ldots, x_n)$. The inputs required with this pattern of production can compactly be written in matrix notation as $\mathbf{A}\boldsymbol{x}$, while the total labour required is $\boldsymbol{l}'\boldsymbol{x}$.
2. *The determination of values*. The value, λ_j, of commodity j is the sum of the direct labour, l_j, expended in its production, and the indirect labour contained in produced inputs required for its production, $\lambda_1 a_{j1} + \cdots + \lambda_n a_{jn} = \boldsymbol{\lambda}'\boldsymbol{a}_j$, that is $\lambda_j = \boldsymbol{\lambda}'\boldsymbol{a}_j + l_j$. The vector of values of commodities, $\boldsymbol{\lambda}'$, satisfies the equation: $\boldsymbol{\lambda}' = \boldsymbol{\lambda}'\mathbf{A} + \boldsymbol{l}'$. It can be written as:

$$\lambda' = \boldsymbol{l}'(\mathrm{I} - \mathrm{A})$$

The value of the net product $\boldsymbol{y} = (\mathrm{I} - \mathrm{A})\boldsymbol{x}$, is equal to the total labour time expended: $\lambda' \mathrm{y} = \boldsymbol{l}'\boldsymbol{x}$. It is the sum of variable capital (wages paid), and total surplus-value. We denote τ as the rate of surplus-value, and v, the value of one unit of labour power, or the share of wages in the net product. These two variables are linked by the relationship $v = 1/(1 + \tau)$.

3. *The example of the table*. Each element in the table (upper-case notation) refers to industries, that is the product of unit variables (lower-case notation) by levels of operation (industries are marked by the subscript j, while vectors have no subscript). Below we will use the notation, P_j, for the price of the output of industry j, p_j for the price of one unit of commodity j, and $\boldsymbol{p}'$ for the vector of unit prices.

$$\begin{aligned}
\text{Constant capitals}: &\quad C_j = \boldsymbol{\lambda}'\boldsymbol{a_j}x_j \text{ and } C = \boldsymbol{\lambda}'\mathbf{A}\mathbf{x}.\\
\text{Variable capitals}: &\quad V_j = v\, l_j x_j \text{ and } V = v\mathrm{l}'\mathrm{x}, \text{with } v = 1/(1+\tau) \text{ or } \tau = (1-v)/v.\\
\text{Total capitals}: &\quad K_j = C_j + V_j \text{ and } K = C + V.\\
\text{Surplus} - \text{values}: &\quad S_j = \tau V_j = (1-v)l_j x_j \text{ and } S = \tau V = (1-v)\boldsymbol{l}'\boldsymbol{x}.\\
\text{Values of commodities}: &\quad \Lambda_j = K_j + S_j = \left(\boldsymbol{\lambda}'\boldsymbol{a_j} + l_j\right)x_j = \lambda_j x_j \text{ and } \Lambda = K + S = (\boldsymbol{\lambda}'\mathbf{A} + \mathbf{l}')\mathrm{x} = \lambda'\mathrm{x}.
\end{aligned}$$

Marx determines the total surplus-value, S, and allocates it proportionally to total capital in each industry, so that the profit rates, r_j, in each industry is uniform: $r = S/K$ (or, equivalently, $1 + r = \Lambda/K$). Profits in each industry are: $\Pi_j = r\, K_j$. By construction, total profits are equal to total surplus-value. The price of production of the total output of industry j is: $P_j = K_j + \Pi_j = (1 + r)K_j$. For the price of one unit of commodity j, one has:

$$\begin{aligned}
p_j &= (1+r)\left(\boldsymbol{\lambda}'\boldsymbol{a_j} + v l_j\right) \text{ and } \boldsymbol{p}'\\
&= (1+r)(\boldsymbol{\lambda}'\mathbf{A} + v\boldsymbol{l}').
\end{aligned}$$

As is obvious, the two equations Sum of values ($\Lambda = \boldsymbol{\lambda}'\mathbf{x}$)= Sum of 'prices of production' ($\mathrm{P} = \boldsymbol{p}'\boldsymbol{x}$) and Sum of surplus-value (S) = Sum of profits ($\Pi = r\, K$) are satisfied.

4. *The determination of prices of production*. In the above calculation, Marx simply transfers the values of inputs to the price of production system instead of estimating them at their prices of production. Prices of production are a stationary price system (in which inputs have the same prices as outputs, as would be the case in a long-period equilibrium) at which profit rates in all sectors are equal to a given r, when the wage is paid at the beginning of the production period:

$$\begin{aligned}
&\boldsymbol{p}' = (1+r)(\boldsymbol{p}'\boldsymbol{A} + w\boldsymbol{l}'), \text{ which implies } \mathrm{p}'[r,w]\\
&= w(1+r)\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}.
\end{aligned}$$

The profit rate equalization conditions are n equations (one for each produced commodity) in $n + 2$ variables, the n prices $\boldsymbol{p}'$, r, and w. Since the accounting units in which prices and the wage are expressed are arbitrary, it is possible without loss of generality to add one further equation normalizing prices, such as $\boldsymbol{p}'\boldsymbol{N} = 1$, where $\boldsymbol{N}$ is a nonnegative bundle of commodities chosen as numéraire for the price system, or, alternatively $w = 1$, which specifies the unit wage as the numéraire.

In the treatment of the transformation problem the most intuitive normalization is to express prices in labour time units. These prices are often called 'direct prices', and the general price level in this metric is determined by: $\boldsymbol{p}'\boldsymbol{y} = \boldsymbol{l}'\boldsymbol{x}$. The price of the net product $\boldsymbol{p}'\boldsymbol{y}$, evaluated at direct prices, is equal to the total labour time expended: $\boldsymbol{l}'\boldsymbol{x}$. This is equivalent to saying that the numéraire is the net

product divided by the total number of hours expended: $\boldsymbol{N} = \boldsymbol{y}/\boldsymbol{l}'\boldsymbol{x}$. Using this numéraire one has:

$$\boldsymbol{p}'[r] = \frac{\boldsymbol{l}'x}{\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}y}\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}$$

Using this relationship and the expression of $\boldsymbol{p}'[r,w]$ above, one can determine the negative relation between wages and the profit rate, à la Ricardo and Sraffa:

$$w = \frac{1}{1+r}\,\frac{\boldsymbol{l}'x}{\boldsymbol{l}'(\mathbf{I} - (1+r)\mathbf{A})^{-1}y}$$

When the profit rate is 0, we have $w = 1$, and $\boldsymbol{p}' = \boldsymbol{l}'(\mathbf{I} - \mathbf{A})^{-1} = \boldsymbol{\lambda}'$: direct prices are equal to values.

5. *The historical transformation controversy.* The dual-system critique is based on comparing the aggregates (sum of values to sum of prices, and sum of surplus-values with sum of prices) under the assumption of a given real wage as a bundle, $\boldsymbol{d}$, of commodities. Thus, the value of labour power and surplus-value are respectively: $v = \boldsymbol{\lambda}'\,\mathbf{d}$, and $S = (1 - v)\boldsymbol{l}'\,\boldsymbol{x}$. Workers are assumed to buy the same commodities when prices of production prevail, so that $w = \boldsymbol{p}'\,\boldsymbol{d}$. Substituting $\boldsymbol{p}'[r,\ w]$, as above, for $\boldsymbol{p}'$ in this expression, the profit rate is the solution of the following implicit equation:

$$(1+r)\boldsymbol{l}(\mathbf{I} - (1+r)\mathbf{A})^{-1}\boldsymbol{d} = 1.$$

One can then calculate Π, which has no reason to be equal to S: in the general case, the second Marxian equation does not hold.

6. *The PNP-UPP LTV.* In the PNP-UPP LTV interpretation, in contrast, the same situation of distribution means the same rate of surplus-value. In general this means that workers will not be able to buy the same bundle of commodities at prices of production. The rate of surplus-values is: $\tau^{\mathrm{p}} = \Pi/W$. If, in the two systems, the price of production of the net product is set equal to its value, of which it is the price form (or, equivalently, if the monetary expression of value is set to 1), that is $\boldsymbol{p}'\boldsymbol{y} = \boldsymbol{\lambda}'\,\mathbf{y} = \boldsymbol{l}'\,\boldsymbol{x}$, then the total price of profits is equal to the sum of surplus-value, of which it is the price form. Thus the two Marxian equations (the first interpreted in terms of the net product) hold.

See Also

- Absolute and Exchangeable Value
- Classical Production Theories
- Competition, Classical
- Labour Theory of Value
- Linear Models
- Market Price
- Marxian Value Analysis
- Marx's Analysis of Capitalist Production
- Natural Price
- Neo-Ricardian Economics

Bibliography

Duménil, G. 1980. *De la valeur aux prix de production*. Paris: Economica.

Duménil, G. 1983. Beyond the transformation riddle: A labor theory of value. *Science and Society* 47: 427–450.

Duménil, G. 1984. The so-called 'transformation problem' revisited: A brief comment. *Journal of Economic Theory* 33: 340–348.

Duménil, G., and D. Lévy. 1984. The unifying formalism of domination: Value, price, distribution and growth in joint production. *Zeitschrift für Nationalökonomie* 44: 349–371.

Foley, D.K. 1982. The value of money, the value of labor power, and the Marxian transformation problem. *Review of Radical Political Economics* 14(2): 37–47.

Foley, D.K. 2000. Recent developments in the labor theory of value. *Review of Radical Political Economy* 32(1): 1–39.

Freeman, A., and G. Carchedi, ed. 1996. *Marx and non-equilibrium economics*. Brookfield, Vermont: Edward Elgar.

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 26: 291–325.

Lipietz, A. 1982. The 'so-called transformation problem' revisited. *Journal of Economic Theory* 26: 59–88.

Marx, K. 1981. *Capital, vols. 1, 2, and 2*. New York: Random House.

Morishima, M. 1973. *Marx's economics*. Cambridge: Cambridge University Press.
Moseley, F. 2000. The new solution to the transformation problem: A sympathetic critique. *Review of Radical Political Economics* 32: 282–316.
Samuelson, P.A. 1971. Understanding the Marxian notion of exploitation: A summary of the so-called transformation problem between Marxian values and competitive prices. *Journal of Economic Literature* 9: 399–431.
Sraffa, P. 1960. *Production of commodities by means of commodities: Prelude to a critique of economic theory.* Cambridge: Cambridge University Press.
Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.
von Bortkiewicz, L. 1952. Value and price in the Marxian system. *International Economic Papers* 1952(2): 5–60.
von Böhm-Bawerk, E. 1890. *Capital and interest*, 1957. New York: Kelley and Millman.

Marxian Value Analysis

John E. Roemer

JEL Classifications
B1

For Marx, the labour theory of value was not a theory of price, but a method for measuring the exploitation of labour. The exploitation of labour, in turn, was important for explaining the production of a surplus in a capitalist economy. In a feudal economy, the emergence of a net product, surplus to the consumption of producers and to the inputs consumed in production, was palpable. For the serf reproduced himself on his family plot of land during part of the week, and then worked for the lord, doing demesne or corvée labour during the other part. There was a temporal and physical division between production for subsistence or reproduction, and production which generated an economic surplus and was appropriated by the lord. Under capitalism, with the division of labour, such a demarcation no longer existed. If capitalism is characterized by competitive markets, where each factor is paid its true 'value', and no one makes a windfall profit by cheating his partner in exchange, how could a surplus emerge? In what manner could a sequence of equal exchanges transform an initial set of inputs into a larger quantity of outputs, with the surplus being appropriated systematically by one class, the capitalists? Marx's project was to explain the origin of profits in a perfectly competitive model, where each factor, including labour, received its competitive price in exchange.

Marx thought he had discovered the answer to this apparent economic sleight of hand by tracing what happened to labour as it passed from the workers who expended it, to the products in which it became embodied, and eventually to the profits of capitalists who sold these commodities. In some of his writings, notably in *Capital*, Volume I, he simplified the argument by assuming that the prices of goods were equal to the amounts of labour they embodied. The embodied labour in a good is the amount of labour necessary to produce that good, and to reproduce all inputs used up in its production. (Assume the only non-produced input is labour.) In particular, this is true also for the good 'labour power'; the embodied labour in a week's labour supplied by a worker is the amount of labour necessary to produce the goods which that worker consumes to reproduce himself for work the following week. If all goods exchange at their embodied labour values (the simplifying assumption) then, in particular, the worker receives a wage in consumption commodities (say, corn) which is just necessary to reproduce himself (which includes the reproduction of the working-class family). The secret of accumulation, for Marx, lay in the discovery that the embodied labour value of one week's labour was, let us say, four days of labour. In four days of socially expended labour, given the existing technology and stock of capital, the consumption commodities necessary to reproduce the worker could be produced. Thus the worker was paid an amount of corn which required four days to produce, his wage for seven days' labour. The surplus labour of three days became embodied in

commodities which were the rightful property of the capitalist who hired the worker. Why would the worker agree to such a deal? Because he had no access to the means of production necessary for producing his consumption goods on any better terms. Those means of production were owned by the capitalist class. (Although the simplifying assumption, that equilibrium prices are equal to or proportional to embodied labour values, is rarely true, Marx conjectured that the deviation of prices from labour values was not crucial to understanding the origin of profits. On this point he was correct. Much ink has been spent on the 'transformation problem', which tries to relate embodied labour values to equilibrium prices in general. As will be shown below, prices need not be proportional to embodied labour values for the theory of class and exploitation to be sensible. Hence the study of the transformation problem is a pointless detour.)

Imagine a corn economy, where there are two technologies for producing corn, a Farm and a Factory: – Farm: 3 days' labour produces 1 corn output – Factory: 1 days' labour + 1 corn (seed) produces 2 corn output. On the Farm, corn is produced from labour alone, perhaps by cultivating wild corn on marginal land. In the capital-intensive Factory technology, seed corn is used as capital. One unit of seed capital reproduces itself and produces one additional corn output with one day of labour. Suppose both techniques require one week for the corn to grow to maturity. Let there be 1000 agents, ten of whom each own 50 units of seed corn. The other 990 peasants own only their labour power. Suppose a person requires one corn per week to survive; his preferences are to consume that amount, and then to take leisure. Assume that if he owns a stock of seed corn, he is not willing to run it down: he must replenish the inputs which he uses up before consuming. What is an equilibrium for this economy, which is guaranteed to reproduce the stocks with which it begins?

Since there are only 500 bushels of seed corn, the required consumption of 1000 corn cannot be reproduced using only the Factory technology, since the seed capital of 500 must be replaced. Capital is scarce relative to the labour which is available for it to employ. The wage which the 'capitalists', who own the seed corn, will offer at equilibrium to those whom they employ will therefore be bid down to the wage which peasants can earn in the marginal Farm technology: 1/3 corn per day labour. At any higher wage, all peasants will wish to sell their labour to the capitalists, and there is insufficient capital to employ them all. (It is assumed peasants have no preference for life on the Farm over life in the Factory. All they care about is rate at which they can exchange labour for corn.) At the wage of 1/3 corn per day, 500/3 peasants become workers in the Factory, each working for three days, planting three units of seed corn, and earning a wage of one corn. This exhausts the capital stock. The remaining peasants stay on the Farm, and also earn one corn with three days' labour. The ten capitalists each work zero days; altogether, they make a profit of (500–500/3) = 333.3 corn, after paying wages and replenishing their seed stock.

In the Factory technology, the embodied labour value of one corn is one day's labour; that amount of labour produces one corn output and reproduces the seed capital used. But the worker, at equilibrium, must work three days to earn one corn. This is so because he does not own the capital stock required for operating the efficient Factory method. His alternative is to eke out a subsistence of one corn by doing three days' labour on the Farm. The worker is said to be *exploited* if the labour embodied in the wage goods he is paid is less than the labour he expends in production. This is the case here, and it is evidently what makes possible the production of a surplus, in an economy where all agents wish only to work long enough to reproduce themselves (and their capital stock). Note this last statement characterizes, as well, the capitalists: in this story, they get 333 corn profits and expend no labour, a result consistent with their having subsistence preferences, where each desires to work only so long as he must to consume his one corn per week.

Contrast this capitalist economy, where three classes have emerged – capitalists, workers, and peasants – to the following subsistence, peasant economy. Everything is the same as above, except the initial distribution of corn: let each of the 1000 persons own initially 0.5 corn. At equilibrium, each agent will work two days and consume one corn. First, he uses the Factory to turn his 0.5 seed corn into 0.5 corn net output, which costs him 0.5 days of labour; then he must produce another 0.5 corn for consumption, for which he turns to the Farm, where he works for 1.5 days. Each agent consumes one corn with two days' labour, an egalitarian society, which is classless. (There are other ways of arranging the equilibrium in this economy, in which one group of agents hires another group to work up its capital stock, while they, in turn, work on the Farm. But the final allocation of corn and labour is the same as in the equilibrium just described.) There is a fine point here: perhaps one should say, in both economies, that the amount of labour socially embodied in one corn is two days (not one, as written above), for that is what is required to produce society's necessary corn consumption given the capital stock and available techniques. This will not change the verdict that the workers in the capitalist economy are exploited, while no one is exploited in the egalitarian society.

Contrast these two economies, which differ only in the initial distribution of the capital stock. Inequality in the distribution of the means of production gives rise to: (1) the production of a surplus above subsistence needs, or accumulation; (2) exploitation, in the sense that some agents expend more labour than is embodied in the goods they consume and others expend less labour than is embodied in what the consume; and (3) classes of agents, some of whom hire labour, some of whom sell labour, and some of whom work for themselves. The exploitation of labour emerges with the unequal ownership of capital, or the 'separation' of workers from the means of production. The existence of an industrial reserve army (here, the peasantry) who have access to an inferior technology to reproduce themselves explains the equilibration of the wage at a level below that which exhausts the product of labour in the capitalist sector. Moreover, exploitation may be an indicator of an injustice of capitalism. If it does not seem fair that a serf must work three days a week for the lord perhaps it is not fair either that a wage labourer must expend more labour than is embodied in the wage goods he receives. That verdict, however, is not obvious and requires further analysis. Although the story can be made complicated, these simple models demonstrate the main features of the Marxian theory of labour exploitation.

Class, Exploitation and Wealth

Consider an economy of N agents, with n produced commodities and labour. The input–output matrix which specifies the linear technology is A, and the row vector of direct labour inputs needed to operate the technology is L. Agent i has an initial endowment vector of goods w^i and one unit of labour power. For simplicity, assume as above subsistence preferences: each agent wishes to earn enough income to purchase some fixed consumption vector b, and not run down the value of his initial endowment, valued at equilibrium prices. After working enough to earn that amount, he takes leisure. It is clear that each agent will only operate activities, at a given price vector, which generate the maximum rate of profit. Normalize prices by setting the wage at unity. For all activities to operate at equilibrium, the commodity price vector p must satisfy:

$$p = (1 + \pi)(pA + L) \tag{1}$$

Prices p obeying (1) generate a uniform and hence maximal rate of profit π for all activities. (The only activities we observe are the ones reported in A and hence without loss of generality, we may assume the profit rate must be equalized for all sectors of production, since agents only operate maximal profit rate activities.)

The vector of embodied labour values in commodities is Λ:

$$\Lambda = \Lambda(1 - A)^{-1} \tag{2}$$

A worker, whose initial endowments are none except his labour power, must earn wages sufficient to purchase the subsistence vector b, which requires:

$$pb = 1 \tag{3}$$

From these three equations, it can be demonstrated (see Morishima 1973; Roemer 1981) that:

$$\pi > 0 \text{ if and only if } \Lambda b < 1 \tag{4}$$

Equivalence (4) was coined by Morishima the 'fundamental Marxian theorem', as it shows that profits are positive precisely when labour is exploited (for the second inequality says that the labour embodied in the wage bundle is less than one unit of labour).

An agent in this model minimizes the labour he expends subject to earning revenues sufficient to buying his consumption b, and to replace the finance capital he uses. Suppose, for simplicity, there is no borrowing and all production must be financed from initial wealth. In general, an agent will optimize by hiring some labour, selling some of his own labour, and/or working on his own capital stock. Let x^i be the vector of activity levels which agent i operates himself, financed with his wealth; let y^i be the vector of activity levels he hires others to operate, which he finances; let z^i be the amount of labour he sells to other operators. His problem is to choose vectors x^i, y^i, and z^i to:

$$\min Lx^i + z^i$$

subject to

(i) $pAx^i + pAy^i \leq pw^i$

(ii) $p(q - A)x^i + p(q - A)y^i - Ly^i + z^i \geq pb$

The first constraint requires him to finance the activities operated out of his endowment, and the second requires that his revenues, net of wages paid and replacement costs, suffice to purchase the consumption bundle b. As well as the price vector satisfying (1), equilibrium requires that the markets for production inputs, consumption goods, and labour must clear. It can be proved that at such a 'reproducible solution', society is divided into five classes of agents, characterized by their relation to the hiring or selling of labour, as follows. There is a class of *pure capitalists,* who only hire labour (y^i is non-zero, but x^i and z^i are zero vectors); there is a class of *mixed capitalists*, who hire labour and work for themselves as well ($y^i \neq 0 \neq x^i$, $z^i = 0$); there is a class of *petty bourgeoisie,* who only work for themselves, and neither hire nor sell labour ($x^i \neq 0$; $y^i = 0 = z^i$); there is a class of *mixed proletarians,* who work for themselves part-time, and also sell their labour power on the market ($x^i \neq 0 \neq z^i$, $y^i = 0$); and there are *proletarians,* who only sell their labour power ($z^i \neq 0$, $x^i = 0 = y^i$). It is clear, from consulting the agent's programme, that this last class comprises those agents who own nothing but their labour power. More generally, the *Class-Wealth Correspondence Theorem* states that the five classes named, in that order, list agents in descending order of wealth. This verifies an intuition of classical Marxism.

There is, as well, a relation of class to exploitation. The *Class–Exploitation Correspondence Principle* states that the agents who hire labour are exploiters and the agents who sell labour are exploited. The exploitation status of agents in the petty bourgeoisie is ambiguous. Exploitation is defined as before: an agent is exploited if he expends more labour than is embodied in the vector b, and he is an exploiter if he expends less labour than that. It is important to note that this relationship of class to exploitation is a theorem of the model, not a postulate. Both the class and exploitation status of an agent emerge in the model as a consequence of optimizing behaviour, determined by the initial distribution of endowments, technology and preferences. These aspects

of agents which in classical Marxism were taken as given (their class and exploitation status) are here proved to emerge as part of the description of agents in equilibrium, from initial given data of a more fundamental sort (endowments, etc.). For this reason, the model described provides microfoundations for classical Marxian descriptions. Generalizations and discussion of the model are pursued in Roemer (1982, 1985a). See Wright (1985) and Bardhan (1984, ch. 13) for empirical applications. For a general evaluation of the Marxian theory of exploitation and class, see Elster (1985, ch. 2, 4 and 5).

From the viewpoint of modern capitalism, many criticisms can be levelled against these stories. Foremost among them, perhaps, is the assumption of subsistence preferences. What happens if agents have more general preferences for income and leisure? The Class–Exploitation Correspondence Principle continues to hold, but the correspondence between class and wealth may fail. It fails, however, only for preference orderings which are unusual: the Class–Wealth Correspondence is true if the elasticity of labour supplied by the population viewed cross-sectionally with respect to its wealth is less than or equal to unity. There can, therefore, be no general claim that exploitation corresponds to wealth, in the classical way – that the poor are exploited by the rich. Whether the exploitation–wealth correspondence holds depends on the labour supply behaviour of agents as their wealth changes.

Exploitation as a Statistic

Note that the fundamental conclusions of classical Marxian value analysis – the association of exploitation with class, in a certain way, and the association of exploitation with profits and accumulation – hold even when equilibrium prices are not proportional to labour values. For the prices of equation (1) are not, except in a singular case, proportional to the labour values of equation (2). Therefore, the usefulness of exploitation theory need not rest upon the false labour theory of value. It is for this reason that the transformation problem, for so long a central concern in Marxian economics, is unimportant.

That usefulness, instead, depends on how good a statistic exploitation is for the phenomena it purports to represent. Does the exploitation of labour explain accumulation? The 'fundamental Marxian theorem' would seem to say so. But, in fact, it can be shown that in an economy capable of producing a surplus, every commodity can be viewed as exploited, not just labour power. If corn is chosen as the value numeraire, then the amount of corn value embodied in a unit of corn is less than one unit of corn, so long as profits are positive. Thus labour power is not unique, as Marx thought, in regard to its potential for being exploited, and it is a false inference that the exploitation of labour 'explains' profits any more than the exploitation of corn or steel or land does. (For versions of this 'generalized commodity exploitation theorem', see Vegara (1979), Bowles and Gintis (1981), Samuelson (1982), and Roemer (1982).)

Is exploitation a good statistic for the injustice of capitalist appropriation of the surplus? Only if the initial distribution of endowments, which gives rise to such appropriation, is unjust. Marx claimed this was so, by arguing that initial capitalist property was established by plunder and enclosure (*Capital,* volume 1, Part 8). But suppose there were a clean capitalism, in which initial inequalities in the ownership of capital were generated by differential hard work, skills, risk-taking postures, and perhaps luck of the agents. Would the ensuing class structure, exploitation and differential wealth indicate an injustice, or would it reflect the consequences of persons exercising traits which are rightfully theirs, and from which they deserve differentially to benefit? These topics are pursued in Cohen (1979) and Roemer (1985b).

In sum, the Marxian theory of exploitation is liberated from the labour theory of value. The link between class and exploitation is robust; but Marx's claim that the exploitation of labour is the unique explanans of accumulation is false. If one's class, defined above as one's relation to

hiring or selling of labour, is important sociologically in determining behaviour (such as collective action against another class) and preferences, then the positive theory of class determination described is of use. Exploitation remains a statistic, of some value, for the inequality in the distribution of productive assets. But in this role, exploitation may not correspond to wealth as in the classical story: if the labour supplied by agents responds with excessive enthusiasm to increases in their wealth, then the rich can be 'exploited' by the poor. The ethical conclusion from an observation of exploitation is in this case unclear.

Even aside from this peculiar case, exploitation is a circuitous proxy for differential wealth in productive assets, and one's normative evaluation of exploitation depends on one's view of the process that generates that inequality. If agents are the rightful owners of their alienable means of production, because they accumulated them based on the exercise of their rightfully owned talents and preferences then exploitation does not represent unjust expropriation. If agents are not entitled to own alienable productive assets, either because they have no right to their talents and preferences (whose distribution is morally arbitrary), or because they came to possess those assets in some other unjustifiable way, then exploitation represents an expropriation. Inheritance, for example, might be an unjust way of acquiring assets which were originally acquired in an untainted manner. The essential question which lies behind the theory of exploitation concerns the fairness of a system of property allowing private ownership of alienable productive assets. The concept of exploitation based on the calculation of surplus labour accounts is, in this writer's view, a circuitous route towards the discussion of that central issue.

Ethical views concerning what kinds of asset may justifiably be privately appropriated change through history. Property in other persons, as in slavery, or more limited rights over the powers of other persons, as in feudalism, are no longer viewed as legitimate. The Marxian theory of exploitation is associated with a call for the abolition of private property in the productive assets external to persons. (Marx himself did not explicitly base his call for the abolition of such property on grounds of fairness, but on grounds of efficiency, despite the clear ethical tone of his attacks on capitalism. For an evaluation of the debate surrounding this question, see Geras (1985).) The cogency of that call must be established independently of the theory of exploitation.

See Also

- ▶ Exploitation
- ▶ Labour Theory of Value
- ▶ Market Price
- ▶ Marx, Karl Heinrich (1818–1883)

Bibliography

Bardhan, P. 1984. *Land, labor and rural poverty: Essays in development economics*. New York: Columbia University Press.

Bowles, S., and H. Gintis. 1981. Structure and practice in the labor theory of value. *Review of Radical Political Economics* 12 (4): 1–26.

Cohen, G.A. 1979. The labor theory of value and the concept of exploitation. *Philosophy and Public Affairs* 8 (4): 338–360.

Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.

Geras, N. 1985. The controversy about Marx and justice. *New Left Review* 150: 47–85.

Morishima, M. 1973. *Marx's economics*. Cambridge: Cambridge University Press.

Roemer, J.E. 1981. *Analytical foundations of Marxian economic theory*. Cambridge: Cambridge University Press.

Roemer, J.E. 1982. *A general theory of exploitation and class*. Cambridge, MA: Harvard University Press.

Roemer, J.E. 1985a. *Value, exploitation, and class*. London: Harwood Academic Publishers.

Roemer, J.E. 1985b. Should Marxists be interested in exploitation? *Philosophy and Public Affairs* 14 (1): 30–65.

Samuelson, P. 1982. The normative and positivistic inferiority of Marx's values paradigm. *Southern Economic Journal* 49 (1): 11–18.

Vegara, J.M. 1979. *Economia politica y modelos multisectoriales*. Madrid: Editorial Tecnos.

Wright, E.O. 1985. *Classes*. London: New Left Books.

Marxism

Andrew Arato

The term 'Marxism' is much overused today: the category is deemed applicable by all sides of political divides unable to agree on anything else. No taxonomic sense, however, can be given to the conceptual chaos behind the wide variety of identifications. Only the historical reasons can be explored in the present context. Here *Marxism* will signify a *tradition* combining two related, originally nineteenth-century, intellectual complexes: (1) a particular, philosophically materialist, comprehensive world view seeking to give a unified, this-worldly explanation to all dimensions of human existence and (2) a 'theory of movement' (R. Koselleck) oriented to the struggles of the industrial working class designed to accelerate historical time, to help bring a (logically, normatively or historically) necessary future closer to the present by 'linking theory and practice'. Both of these complexes are derived from the philosophy of history (or one of the philosophies of history) of Karl Marx, but the founder had little interest in working out a general *Weltanschauung*. In this respect Friedrich Engels was, in works such as *Herr Eugen Dühring's Revolution in Science* and the posthumous *Dialectics of Nature*, the founder of Marxism. Those who look back to the original work of Marx as against the tradition founded by Engels should be identified by the adjective '*Marxian*', as in Marxian philosophy, economics, social theory or anthropology, etc. Nevertheless Marx's relation to Marxism is too complex to allow a neat division between the two. As Lukacs first demonstrated in 1923, Engels's interpretation of Marx's oeuvre in the sense of a generalized worldview and a unified science missed the actual philosophical depth of the latter's theory of history, social theory and critique of political economy. The cost was the elimination, misunderstanding or de-emphasis of fundamental concepts like alienation, reification, fetishism, praxis, subject, etc. Nevertheless, the great power and influence of Engels's synthesis came from Marx's own marriage of science and philosophy of history, bringing together the intellectual prestige of enlightenment with the motivating power of the concepts of romanticism. In another respect as well, Marx, despite having supposedly declared that he was not a 'Marxist', contributed to the foundations of Marxism. He *did* interpret his own thought in all of its phases as providing a theory of movement based on a philosophy of history whose major concepts included (typically) historical stage, transition, revolution and progress. The specific content of this theory was meant to be both an interpretation of *the meaning* of the movement of the industrial working class, and contribution to its enlightenment. No doubt, Marx understood scientific commun*ism* or social*ism* not only as the diagnosis of the crisis of this time, but also as its resolution in anticipation and acceleration of a desired future. This future was conceived in different ways in his various works, but always involved the abolition of differentiated economic and political institutions and the creation of conscious, planned, collective control over economic life as well as direct, democratic participation in all 'political' processes. It is important to note on the one hand that Marx's views of the transition to such a condition were heterogeneous and at different times involved authoritarian étatistic forms (*Communist Manifesto*), the direct democratic (*Civil Wars in France*) and even parliamentary democratic forms (various addresses, and possibly *Class Struggles in France* as well as *The 18th Brumaire of Louis Bonaparte*). Common to all these forms on the other hand was the postulate of the abolition of the division of state and civil society, i.e. an independent civil society with its mediating institutions. The plurality of forms of transition worked out by Marx points to the different politics of later Marxisms, the underlying hostility between civil society and the state strengthening the logic of all the politically significant varieties.

The historical influence of Marxism had been nothing short of spectacular. Until World War II and in some countries until the 1970s, it was the dominant ideology of the various European continental labour movements in Social Democratic,

Communist, Socialist and Euro-Communist forms. The theoretical works oriented to these movements were often of the highest quality; it is enough to mention only the best works of Kautsky, Bernstein, Hilferding, Luxemburg, the Austro-Marxists, Lukacs, Gramsci and countless others. Secondly, from 1905 or so to as late as the 1970s, another version of Marxism eclipsed even nationalism as the dominant revolutionary ideology of 'underdeveloped' or 'peripheric' agrarian societies. Again significant intellectual output accompanied this process, from Lenin, Parvus, Trotsky and Bukharin to Mao, Guevera and Cabral. Finally, from 1917 but more globally from 1945, an increasing number of regimes have used a version of Marxism (Marxism-Leninism, Soviet Marxism) as their 'science of legitimation', their official state cult. While the early phases of this process even here involved serious intellectual work – for example, the 1920s soviet debates about economic development (Preobrazhensky, Bukharin and others) and the problems of law and politics (Pashukanis, Stukha, et al.) – from the 1930s Marxism in power always meant tremendous simplification and even falsification of the doctrines of classical Marxism (not to speak of Marxian theory). However, this intellectual reduction was for a time amply compensated by the prestige of successful revolutions. Thus Communist movements outside the powers of Communist regimes continued to attract an astonishing number of philosophers, scientists, economist, historians, social theorists, legal scholars, writers, poets and plastic artists, even as their counterparts were suppressed in the Soviet Union, and later Eastern Europe and China.

Today one senses an ever deepening exhaustion of Marxism in all of the areas of its greatest historical influence. Among the mass parties and unions of European labour only an ever smaller minority remains or even calls itself Marxist. The official Marxisms of the established regimes are increasingly ritualized, the operative beliefs of the leaders and ideologists themselves have been shifting toward other doctrines: nationalism, authoritarian technocracy, pragmatism, great power politics, small nation *raison d'état*. Even among third world movements, the remaining area of dynamic influence, Marxism today has more powerful competitors than ever before.

The idea of the 'crisis of Marxism' is almost as old as Marxism itself. There are, nevertheless, deep-seated reasons today why the epoch of Marxism as defined here (and not the rich and varied influence of the thinker Karl Marx) is over. As a world-view Marxism has certainly been shaken by the general secularization, decentralization, differentiation and pluralization of world-views that is for better or worse the hallmark of modernity. More importantly, as a theory of movement it is paradoxically the very successes of Marxism that have undermined it. It is this second aspect that is decisive, because inhibiting the often attempted transformation of Marxism in a direction no longer bound to a comprehensive, metaphysical worldview, i.e. toward a 'critical theory' or a 'philosophy of praxis' in the sense of the early theorists of Western Marxism: Lukacs, Korsch, Gramsci, Horkheimer and Marcuse.

In most general terms then, politically failed attempts can be described as breaks with the tradition of Engels in order to build a new *Marxism* around the *Marxian* legacy itself. This implied in each case not only a return to the Hegelian foundations of Marx's thought, a revival of the key concepts of alienation, reification, consciousness of subjectivity, but also a primary emphasis on the theory of movement dimension and in particular the mediation of theory and praxis. The latter emphasis however depended on an intellectually adequate and politically favourable response on the part of those to be addressed by theory, an impossibility in the case of the actually successful regimes and movements. In the case of the Soviet regime it hardly mattered that the very first Western Marxists sought to give it a new philosophical justification. The abandonment of the metaphysical world-view of Marxism could not be contemplated first of all because Lenin in his *Materialism and Empiriocriticism* helped to canonize it. More fundamentally, such a world view, along with the deterministic and closed structure it lent to Marxian philosophy of history, allowed the ritualization of the doctrine as a new state cult. The anti-authoritarian biases of Western Marxism, formulated in a much repeated critique of bureaucracy,

expressed well the incompatibility of Western and Soviet Marxism, whatever the particular political choices of individual Western Marxists. In the case of Social Democracy the problem was not so much the abandonment of the general world-view of Engels et al., for the Austro-Marxists and other Kantian socialists did this *within* the existing organizational frameworks. The renewed stress on a theory of movement emphasizing revolutionary rupture and future orientation was understood by Social Democrats (except the Austro-Marxists) as expressing the spirit of the rival Bolshevism, with which some founders of Western Marxism were associated. Again more fundamentally the clash was between present and future orientation, bureaucratic organization and movement, (welfare) statism and anti-statism.

In world-historical terms Marxism represented a set of ideological and political responses to the epoch of classical capitalism, to the first stage of K. Polanyi's 'Great Transformation': the self-regulating market. Both of the two major types of outcomes with which 'successful' Marxist movements had an 'elective affinity' were powerfully state strengthening: the emergence of étatist forms of modernization-industrialization where private capital could not or could no longer promote economic development (A. Gerschenkron) and the construction of democratic, interventionist, welfare states in already developed capitalist countries where the 'normal' operation of the self-regulating market produced disastrous consequences not only for the substratum of human life but also for the market economy itself (K. Polanyi). In both cases versions of Marxism were dynamic and influential as ideologies of the process ('revolution' and 'reform') and were made gradually irrelevent by the results. The issue was not only that soviet-type societies and democratic welfare states are not the Marxist 'utopia' or even the 'transitional society'. More damaging was the fact that the Marxian philosophy of history in any of its original versions had no place for a new, industrial form of domination; neither capitalism or socialism, or even a hybrid of the two, while the Marxian critique of political economy had no concepts to deal with the 'primacy of the political' and tended to exclude the possibility of a reconstructed, capitalist society involving a good deal of state interventionism and redistributive activity built upon the institutionalization and integration of the working class and class-based conflict.

Neither the formidable attempts of Trotsky (*Revolution Betrayed*, etc.) and of historians to depict the Soviet Union as a deformed workers' state, nor bureaucratic collectivism or state capitalism, nor the various theories of state-organized monopoly or state monopoly capitalism could successfully address the new contexts. The reason was of course that all these attempts involved a desperate desire to stay within the historico-philosophical framework of Marxism that was deeply enmeshed within the ideological counter-attack of the modern state, or more properly of state-strengthening elites, against the apparently more powerful (under classical capitalism) institutional complex of the modern economy. For this reason above all Marxism has found it hard to remain or to become a critical theory where a version of the modern bureaucratic state became the centre of societal steering and control (Soviet-type societies) or even where modern state and capitalist economy shared steering and control functions in historically unprecedented combinations (welfare states).

The failure of Marxism in face of the modern state has been manifested most openly in the context of the emergence of new types of social movements. The problem was not simply that movements have now been forced to oppose the state (something hardly unprecedented) but, rather the very goals were now reconceived as 'society strengthening'. As a result of important historical learning experiences victory was no longer seen in terms of inclusion in state power ('Reform') or even as smashing the state ('Revolution') but, in the case of the most advanced segments of movements, as rebuilding civil society and controlling (rather than abolishing) market economy and bureaucratic state. Unlike the Social Democracy the state was found not to be a neutral force that could be simply occupied and used by different classes. But as against Bolshevism and even the older Western Marxism (both more orthodox here than Social Democracy) the

programme of smashing the state and the utopia of the withering away of the state were now implicitly recognized as powerfully étatistic. If the modern state does not simply express the power of a class in society but of an independent structure, then contrary to the claim of Engels in *Anti-Dühring* the project of the withering away of the state by way of the abolition of classes and the nationalization of the means of production cannot be successful. On the contrary the very attempt presupposes enormous concentration of state power feeding on the continued social division which it itself constitutes. The actual experience of Marxist revolutionary states (as all previous revolutions according to the judgement of Marx in *The 18th Brumaire*) was a dramatic confirmation of this process, the results representing the most serious challenge to all who seek to defend Marxism in the face of the projects of the new movements.

Of course we cannot yet speak of the actual death of Marxism. In Soviet-type societies as well as in various third world adaptations of this model, Marxism-Leninism still exist as the official state doctrine and cult. However, in a period of the crisis of this model and the failure (after 1968) of bloc-wid reform strategies, the rational elements of any Marxism (e.g. the project and the expectation of dysfunction and crisis free incremental economic development) had to be eliminated. The result is a ritualized, de-intellectualized doctrine increasingly cynically held. This affects Third World contexts, where under Western hegemony and/or right-wing authoritarian domination something like earlier revolutionary versions of Leninism are still upheld. Here the future orientation of Marxian theory is increasingly determined by the actual outcome of the Soviet model which is known and which is decreasingly attractive as against a utopia drawn from Western sources, presupposing Western tradition, one that was nowhere realized. Third World Marxism increasingly reduces to a merely present orientation that involves primarily the assumption of a specific position in world conflicts, or (less attractively) to a conscious preparation for future power positions of the Soviet type. Since the mid-1970s neither of these orientations seem to be able to match more dynamic and radical ideologies where these are available in particular national self determination and religious fundamentalism. Paradoxically, Marxism in the Third World is at its most influential where it is allied with the cultural and political forces of its old enemies: nationalism and radical Catholicism.

The intellectually most significant attempts to renew Marxism in our time (mid-1950s and after) occurred in its historical homeland: Western and Central Europe. The goals of all the relevant trends were to work out critical theories of Soviet type and/or advanced capitalist socialist societies, for the orientation of new types of opposition. What is common among all of them was the attempt, once again, to break with Marxism as a general, metaphysical world-view, while the dimension of a theory of movement was held on to and built upon. Three (and a possible fourth) stages or types can be distinguished in this whole development – each with a relatively different relationship to actual movements.

(1) *Revisionism* was the first of those, primarily of East Central European origins, but radiating to the communist parties of the West as well as to the Soviet Union. Revisionism involved the recovery of the democratic socialist stress of turn of the century revisionism (E. Bernstein et al.), and the corresponding abandonment of Marxian doctrines deemed especially antidemocratic, e.g. the dictatorship of the proletariat. The political high point of revisionism was the preparation of 1956 in Poland and Hungary. The intellectual foundations of revisionism were, however, rather shallow and eclectic; on the one hand non-objectionable features and even the style of Marxism-Leninism were adhered to (at times for pragmatic reasons) and on the other there were attempts to make the very same set of doctrines vehicles for 'socialist legality', industrial democracy, market socialism and at times party pluralism. Thus the failure of revisionism was not only political but also theoretical, and all subsequent attempts to renew Marxism involved far greater efforts at genuine theory building.

(2) The *Renaissance* of Marxism (Lukacs) otherwise called the philosophy of praxis (partially overlapping with Revisionism) was of simultaneously West European, East European and Yugoslav origins. Its return was not only to Marx's own philosophy and social theory, but also to its real predecessor the Western Marxism of the 1920s and 1930s. However, unlike Western Marxism, the Renaissance of Marxism involved at least some attempts to apply Marxian theory to the critique of Soviet type societies. The Renaissance of Marxism, in spite of its common intellectual style was oriented to different political projects in East and West. In East Europe it became the ideology of the internal democratization of communist parties and of reform from above, culminating in the Czech events in 1968. In the West the relevant political streams were the New Left and in some countries movements of working class youth, culminating in the French 1968 as well as the Italian 'hot autumn' 1 year later. Almost all major trends in the Renaissance of Marxism, like their Western Marxist forerunners, were open to at least some elements of non-Marxist thought: Husserl and Heidegger, Freud and Weber, structural linguistics and anthropology, Keynes and the neo-Keynesians were of major influence. The classical intellectual products included the works of Polish, Yugoslav, Czech and Hungarian praxis philosophers, Modzelewski and Kuron's *Open Letter*, Sartre's *Critique*, the early *Socialisme ou Barbarie*, Baran and Sweezy's *Monopoly Capital*, Marcuse's *Soviet Marxism* and *One-Dimensional Man*, the revivals of the older critical theory in West Germany and the United States, of Gramsci in Italy, and finally a good deal of Marx scholarship in France and West Germany. The purely intellectual achievements of the Renaissance of Marxism were significant; to it we owe the present availability of all dimensions of the ouevre of Marx. Politically, however, the Renaissance of Marxism was doomed when established regimes turned away from internal reform in the East, and when the New Left dissipated or proceeded to imitate authoritarian Marxisms imported from the Third World.

The last two stages of the attempted renewal of Marxism, its *Reconstruction* (3) and *Transcendence* (4) are to be located first of all in the changed contexts of movements: the new social movements of the West and the democratic opposition of the East. Each of the two types of attempts is to be found in both world contexts, but the *Reconstruction* of Marxism has its centre in the West while the *Transcendence* of Marxism is primarily Eastern even if there has been very strong French participation. Interestingly enough the movements addressed by both trends are those that reflectively incorporate and criticize the experience of the étatist response to the capitalist economy – thus in a sense they are all 'post Marxist'. Nevertheless, both the Reconstruction and Transcendence of Marxism seek to address post-Marxist movements in ways residually continuous with the tradition. In the case of the first, associated primarily with younger members of the Frankfurt School (Habermas, Offe, Wellmer et al.) the aim was (at least until the mid-1970s) to serve the normative project of human emancipation inherited from Marx and Western Marxism with entirely new theoretical instruments: linguistic philosophy, hermeneutics, systems theory, symbolic interactionism, structural functionalism, social scientific conflict theory, developmental psychology, etc. The Marxian critique of political economy preserved a certain model character for this trend, but only for a non-economistic crisis theory. In the early 1980s it has become clear that the new movements of ecology, feminism, youth and peace (rather than some intellectual new class as some have charged) were the projected addresses of this theoretical strategy.

The transcendence of Marxism, anticipated by Merleau-Ponty's *Adventures of the Dialectic* is represented by thinkers such as Castoriadis, Lefort, Touraine and Gorz in France, and above all a whole series of East European writers, publicists and philosophers like Kolakowski, Juron, Michnik, Kis, Bence and Vajda. In the United States this position is represented by *Telos*, a

journal of radical social thought. The figures of this intellectual topos are not simply non-Marxists or anti-Marxists: they declare their rejection of both dimensions of Marxism as defined here (and especially their foundation: Marx's philosophy of history) while continuing to rely on some key categories of the tradition (theory and practice, state and civil society). This preservation, however, involves some characteristic twists: in particular the normative project of the radical democratic unification of state and society is rejected in the name of an independent civil society and its mediating institutions.

The specific achievement of the Transcendence of Marxism and of the East European opposition addressed by it is the thematization of a self-limiting radical democracy seeking to rebuild or democratize independent societal institutions, seeking to control rather than to absorb modern state and economy. Such a model also seems to correspond to the project of the non-fundamentalist wings of new social movements in the West (the French CFDT and second left, the *Realpolitiker* fraction of the German Greens, etc.) even if the great rhetorical presence of fundamentalists tends to occlude this fact. Furthermore, the project corresponds also to the programmes of present day democratizing movements under Latin American dictatorships: in particular in Argentina, Brazil and Chile. The increasing universality of self-limiting radical democracy or the democratization of civil society has had, since the mid-1970s an apparently decisive effect even on the theorists of the Reconstruction of Marxism, in particular the work of Habermas in the 1980s.

On the other side the achievement of the Reconstruction of Marxism has been above all the creation of a social theory that has surpassed the best in the Marxsian tradition in complexity, scope and self-reflection. Unfortunately we cannot yet speak of the post-Marxist generation either equalling or fully appropriating this social theory. Thus a synthesis of the normative concerns of the transcendence of Marxism with the analytical power of the works of the Reconstruction of Marxism has occurred more in the West than the East, more in Germany than in France. At the same time the politically most advanced expression of such a synthesis took place in the East, in the Polish democratic movement, and was better understood in France than in Germany. Thus it is the paradox of the present situation of even the illegitimate offsprings of Marxism (reminiscent of one described by Marx in 1843) that there is a geopolitical disjuncture between the most advanced version of theory and the most self-reflective form of political action. It is too early to tell if theory and praxis can be brought closer together and if any version of Marxism can serve as a bridge between them. The archaeological link between all versions of Marxism and the strengthening of the state speaks against such possibility in the epoch of the offensive of different models of civil society against the state. The popular understandings of Marxism cannot be easily liberated from previous experience. It may also be the case that the incorporation of the critique of the state in any reconstructed Marxism is destined to burst all conceivable forms that would guarantee even a tenuous continuity with the tradition.

M

See Also

- ▶ Communism
- ▶ Full Communism
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Socialism
- ▶ Utopias

Bibliography

Arato, A., and P. Breines. 1979. *The young Lukacs and the origins of Western Marxism*. New York: Continuum Press.

Bottomore, G.T., et al. 1983. *A dictionary of Marxist thought*. Oxford: Basil Blackwell.

Cohen, J.L. 1982. *Class and civil society. The limits of Marxian critical theory.* Amherst: University of Massachusetts Press.

Habermas, J. 1976. *Communication and the evolution of society.* Boston: Beacon Press.

Kolakowski, L. 1978. *Main currents of Marxism*. 3 vols. Oxford: Oxford University Press.

Koselleck, R. 1985. *Past futures. On the semantics of historical time*. Cambridge, MA: MIT Press.

Piccone, P. 1983. *Italian Marxism*. Berkeley: University of California Press.
Polanyi, K. 1944. *The great transformation*. New York: Rinehart.

Marxist Economics

Andrew Glyn

By Marxist economics we mean the work of those later economists who based their methodology and approach on the work of Karl Marx. Excluded from discussion here is the enormous body of exegetical literature seeking to amplify the genesis of and development of Marx's own thinking (Rosdolsky 1968). Before discussing three areas where the contribution of Marxists has been most striking and important, it is helpful to bear in mind certain general features of their approach which could be said to separate them off from other traditions in economic theory.

Marxist economists view the capitalist system as essentially *contradictory*, in the sense that its malfunctions derive in an essential way from its structure, rather than representing 'imperfections' in an otherwise harmonious mechanism. At the heart of this structure is the relationship between capital and labour, which is necessarily an exploitative one. The conflict which results has a crucial influence on the way the capitalist system develops in every respect, from the form of technologies developed to the pattern of state policies adopted. Capital accumulation, the motor of the system, cannot therefore be analysed simply in quantitative terms: the structural changes in the economy which it brings are influenced by, and in turn help to shape, relations between the classes. So while the underlying logic of capitalism has remained unchanged, its history can be divided into different periods characterized by particular sets of class relations, technologies, state policies and international structures.

If some of these ideas would seem practically self-evident to economists with any interest in economic history, this underlines the powerful confirmation which the past century has provided for many of Marx's central ideas. It cannot, unfortunately, be said that mainstream economic theory has caught up with this, hiding, under ever more powerful formal techniques, an unchanging conceptual superficiality in its approach.

The body of Marxist economics which underpins the approach of Marxist economists to the analysis of particular phases and aspects of capitalist development may be divided into three main parts: (1) the labour process; (2) value, profits and exploitation; (3) capital accumulation and crises. What follows represents a brief survey of debates around and developments of these aspects of Marx's work; it is necessarily narrowly 'economic' (excluding work on the theory of the state and of classes) and concentrates on theoretical debate rather than on historical application.

The Labour Process

Marx's most fundamental criticism of his Classical predecessors, and especially of Ricardo, was that they failed to analyse how the capitalist system emerged as a specific mode of production resulting from a particular historical process. The dispossession of previously independent producers led to a division of society into workers, with only their labour power to sell, and employers who owned and controlled the means of production. This ownership was the basis of the profit appropriated by the capitalists, for it gave them control over the process of production itself. It allowed the capitalist class as a whole to force the working class to work longer than was required to produce their means of subsistence. Marx paid special attention to this control over the labour process, analysing in great detail how the development of machinery qualitatively increased the depth of this control by literally taking the pace of work out of the hands of the workers. This stress on the process of production as a *labour* process is arguably the most important distinguishing feature of Marxist economics as compared to other schools, which analyse production solely in technical terms (Rowthorn 1980, ch. 1).

It was not, however, until more than 100 years after the publication of volume I of *Capital* that his analysis of capitalist control over the labour process was applied to subsequent developments. Harry Braverman's *Labour and Monopoly Capital* (1972) had as its central theme the striving of employers to separate the conception of tasks from their execution, in order to preserve and enhance their control over the process of work. Frederick Taylor's system of Scientific Management, for example, analysed the operations required of skilled machine tool operatives so that 'scientific' timings could readily be allocated for new types of work. Ford's introduction of the assembly line was similarly intended to force a certain pace of work. Subsequent writers have extended this analysis to describe systems of 'bureaucratic' control exercised in large modern corporations, where effort is secured by payment systems allowing a steady progression of earnings for loyal employees (Edwards 1979).

This more recent work is a revision, as well as an extension, of Marx's own analysis. In his conception of 'modern industry' control over the pace of work was exercised by the machine itself, which carried out the operations on the materials automatically, leaving the worker as a simple machine minder who fed the machine and dealt with minor malfunctions. This pattern, which Marx saw in contemporary developments in the textile industry, has not become the universal one. For in many types of production the worker still carries out operations on the materials. This has made it necessary for the employers to attempt to gain control over the speed of work by mechanical contrivance (the production line which obliges the worker to carry out tasks at a set speed) or organizational means (scientific management). Moreover, it has been more recently argued that 'Fordist' systems of mass production, where there is a minute division of labour, are giving way to more flexible systems where workers perform a greater range of tasks (Aglietta 1979). This reflects the trend towards more sophisticated consumer goods, which demand shorter production runs and more model changes, and also the problems of overcoming the employee dissatisfaction with mindless and repetitive work which exploded in a number of countries at the end of the 1960s.

Marx's fundamental insight remains, however, the inspiration of this whole body of work, focusing on an issue of tremendous contemporary significance as employers struggle with the necessity of restructuring production in the fiercely competitive conditions of the 1980s (see as an example Willman and Winch 1985). Only very recently has mainstream economics begun to address the problem of controlling work, and even then, as argued by Bowles (1985), from a less compelling perspective.

Value, Profits and Exploitation

Critics of Marx, from Böhm-Bawerk (1896) onwards, have always contended that his theory of profits and exploitation was fatally flawed by his reliance on a simplistic 'labour theory of value' – that commodities exchange in proportion to the amount of labour time required to produce them. If the price of a commodity was determined directly by this 'embodied labour', then the wage would directly measure the labour time required to produce the goods which workers bought in order to maintain themselves (the *value of labour power* in Marx's terminology). Profit, being the difference between the value added by the worker and the wage, would similarly measure directly the excess of time worked over the value of labour power, that is the surplus value produced by the worker while under the employer's control. At the level of society as a whole, total profits would be a direct measure of the surplus labour performed by the whole working class, that is the time worked beyond that necessary to reproduce the means of subsistence. Marx's *rate of exploitation*, the ratio of surplus value to the value of labour power, would be directly reflected in the ratio of profits to wages. Marx's insistence that the source of profit was the capitalist's ability to control the labour process, and thus force the working class to perform surplus labour, would receive a clear expression.

Marx himself was quite aware that the assumption he employed in *Capital*, Volume I, that

commodities exchange at their values, that is, in proportion to the labour required to produce them, was a simplification designed to highlight the overall relation between capital and labour. In Volume III he explains that this assumption will only hold when the *organic composition of capital*, that is the ratio between the value of outlays on machinery and materials (*constant capital*) and on wages (the value of labour power), is equal across industries. Where the organic composition differs across industries, then the surplus value produced by workers in a particular industry would represent a greater or lesser rate of profit on total capital employed depending on whether the organic composition was low or high. But exchange in proportion to labour time would inevitably mean that the capitalists within an industry received surplus value equal to that produced by their workers. This is because the commodities they received in exchange would be of equal value to those produced, thus leaving a surplus value for the capitalists, after setting aside what was required to pay for constant and variable capital, just equal to the surplus value their workers produced. Accordingly exchange in proportion to labour time would imply unequal profit rates across sectors, which is impossible under competitive conditions.

Marx's own solution was to propose that commodities exchange not at their values, but at their prices of production. These represented a modification or transformation of values in order to ensure equal rates of profit across sectors despite unequal organic compositions of capital. It was simple for him to show that such prices of production implied that industries with a high organic composition, and which therefore needed to appropriate more surplus value than its workers produced to compensate for the bigger outlays on constant capital, would have to have a higher than average ratio of price of production to value (and vice versa for low organic composition sectors). So Marx's solution to the transformation problem involved a simple redistribution of total surplus value away from labour intensive industries.

As von Bortkiewicz (1906) was the first to point out, Marx's solution to the transformation problem was incorrect. When constructing his prices of production Marx adds the average rate of profit applied to the values of the inputs. But if commodities do not sell at their values then capitalists are not purchasing their inputs at their values but at their prices of production. So correct prices of production have to be calculated on the basis of a simultaneous transformation of inputs *and* outputs from values to prices of production. Marx was actually aware that this further step was necessary but thought, not unreasonably, that it would make no important difference. Unfortunately he was wrong.

For the 'correct' solution to the transformation problem makes it impossible to maintain Marx's equality between such value aggregates as surplus value and the total value of output on the one hand, and their price correlates, profits and total output in money prices. Much subsequent literature (see von Bortkiewicz (1906) and the later generalization by Seton (1957) concentrated on describing the circumstances under which at least one of the 'invariances' between the price and value systems would hold. It can be argued, however, following the Uno school of Japanese Marxists (see Itoh 1980), that this search for numerical equality between surplus value and profit is wholly misconceived, stemming from Marx's failure to maintain consistently his Volume I distinction between the *substance* of value (labour time) and its *form* (money prices). Any attempt to force numerical equality is artificial and thus misleading.

Even so this does not dispose of the 'problem'. For the correct, simultaneous solution also makes the rate of profit on capital employed different from Marx's general rate of profit, calculated as the ratio of surplus value to the value of capital employed (see von Bortkiewicz 1906, and Steedman 1977). What might seem more damaging still is that the rate of exploitation in value terms is not in general equal to the ratio of profits to wages. So Marx's basic expression of the extent of capitalist domination does not find a direct reflection in the money aggregates.

This in fact does not damage Marx's theory at all. The ratio of profits to wages reflects the ratio of surplus product to the bundle of wage goods as manifested in the exchange process (aggregate wages must represent the price of production of

all wage goods and aggregate profits the price of production of the surplus product). The rate of exploitation is the ratio of the work done to produce the two bundles. These two ratios will only be equal when the organic compositions in the sectors producing the wage goods and surplus products are equal. Clearly there is no theoretical necessity for this to hold, though empirical estimates by Woolf (1979) suggest that the deviation of relative prices from relative values for these bundles of commodities may be rather small.

This divergence between the form of exploitation (the ratio of profits to wages) and its real substance (the ratio of surplus value to the value of labour power) can be readily accepted Using Sraffa's construction of a standard commodity to show what pattern of industries would ensure equality between the two ratios seems to add rather little (see Medio 1972). Retreating to the rather grandly named Fundamental Marxian Theorem, that positive profits require positive surplus value (Morishima 1973), also seems unnecessarily defensive in that it fails to explain clearly the relationship between the price and value dimensions. It is important to emphasize that this interpretation of the transformation problem does not establish the case *for* analysis in terms of values. It merely shows how the value categories can be reconciled with the surface phenomena of profits and prices.

Further controversy over the adequacy and usefulness of Marx's theory of value has revolved around two further issues. The whole 'transformation problem' assumed that the values of commodities can be unambiguously defined as the labour time socially necessary for the production of a commodity at prevalent degrees of mechanization, skill, and intensity of work. But critics from Böhm-Bawerk onwards have disputed that differnt types of labour can be 'reduced' to simple labour (see Rowthorn 1980, ch. 6). It has further been argued (Steedman 1977) that in situations of joint production labour values may not be determinable at all. If the output of shepherds is mutton and wool, how can their labour be allocated between the two products? If the employers used the wool and the shepherds ate the mutton it would not be possible to divide the shepherds' total working day into the necessary labour worked to produce the means of subsistence and the surplus labour worked for the employers. More generally, where there are different methods of joint production, the standard method of deriving labour values can lead to their being negative. Negative surplus value has been shown to coexist with positive profits (Steedman 1977), though not uncontroversially (King 1982).

These criticisms have at least made Marxists accept that there are real analytical difficulties in drawing up consistent value schema. The riposte of some (e.g. Himmelweit and Mohun 1981 drawing on the work of I.I. Rubin 1928) that the whole project of deducing values prior to their reflection in market prices in a misguided 'neo-Ricardian' exercise has not found much favour. It seems to abandon any *quantitative* aspect to value theory, leaving simply a *qualitative* emphasis on understanding exchange as an exchange of labours (see Hilferding's reply in Böhm-Bawerk [1896]; Sweezy 1942; Rubin 1928).

The conceptual problems in formalizing value theory hardly differentiate it from other theoretical constructs. The most serious attack on it has come from those claiming that it is *redundant*, that it adds nothing to the conceptualization of equilibrium prices and profits based on physical quantities. This criticism goes back at least to Joan Robinson (1942), was formalized by Samuelson (1971) and re-emphasized by Steedman (1977). Following Sraffa (1960), it is argued that prices and profits can be derived directly from knowledge of the real wage and the requirements of labour and means of production required to produce commodities, and that values can only be derived from the same data. Thus it is said that it is unnecessary to go via values to reach profits (even assuming values can be unambiguously defined). This attack has confronted Marxists with the question – what precisely is it that values are designed to do?

The justifications for using labour as the central conceptual category, and thus analysing exchange and exploitation in terms of embodied labour time have ranged from rather abstract statements of the fundamental role played by labour in Marx's whole theory of society (Shaikh 1981), to the

claim that working with values focuses the analysis on labour's part in production (Dobb 1937). Sen (1978) points out that we naturally focus on the human contribution to production just as we focus on an artist's part in a sculpture. Indeed, critics of value theory never stop to question why they are perfectly happy to regard labour productivity as a vital concern (over time, across countries etc.) but object to the concept of value (which is just the inverse of labour productivity). Certainly for those who accept the central role of the economic surplus produced by the working class in the development of society, and the relations on the factory floor as the key to the production of this surplus, then analysis in terms of labour time is clear and simple. If we want a vivid and forceful way of analysing the relation between capital and labour then labour time seems the obvious category to use. After all what capitalists make workers do is *work*.

Accumulation and Crises

Marx's *Capital* was aimed not only at uncovering the basis of capitalist exploitation but above all at revealing capitalism's 'laws of motion'. Marx argued that competition between capitalists was fought out by their investing in new, more efficient techniques of production and that the economies of scale which this brought acted as a pressure forcing individual capitalists to accumulate (a very different conception from the neoclassical idea of accumulation as trading off present for future consumption – see Marglin 1984). The outcome of this process was the increased concentration of industry (termed centralization by Marx), which was further accelerated by the development of the credit system. Many Marxist writers, from Hilferding (1910) to post-war Marxists (Mandel 1962) have documented this trend, with the conclusion being drawn on occasions that the extent of monopolization was actually destroying the pressure to accumulate (Baran and Sweezy 1966). This seemed to be contradicted, however, by the great boom of the 1950s and 1960s in Europe and Japan, and the spread of international competition which it brought.

For Marx the impact of accumulation, both on the working class and on profits, was dominated by its presumed labour-saving form. Marx argued that higher productivity required an increased volume of constant capital per worker (what later economists have called the capital–labour ratio). While this is not necessarily the case, since new techniques may economize on constant capital, subsequent experience has entirely vindicated Marx's view. What has been more controversial are the implications of this for employment, wages and the rate of profit.

A rising mass of constant capital per worker implies that employment grows more slowly than the capital stock. But whether or not this leads to a rising or falling *reserve army of labour* depends on the strength of accumulation, the rate at which technical progress is labour saving, and the growth of the labour force. In the advanced countries at least, the trend has indeed been for the capitalist sector to overcome pre-capitalist sectors like peasant agriculture, but for those 'set free' to be absorbed into wage labour. It is important here to distinguish the impact of the trend of accumulation on employment (at full utilization of capacity), from periods of 'cyclical' unemployment, which may be of extended duration of course, resulting from the underutilization of capacity during crises. The mass unemployment of the 1970s and 1980s in Europe, for example, is obviously due mainly if not wholly to the crisis of accumulation (that is, the lack of it), rather than to the form accumulation has been taking.

Despite periodic bouts of unemployment there has been a tendency for real wages to grow in line with labour productivity in the advanced countries, that is, for the profit share to be roughly constant over time or even to decline. Despite measurement complications concerning the treatment of self-employment, this suggests that Marx's rate of exploitation has not shown the tendency to increase which he expected would be ensured by the reserve army of labour. Some authors (Gillman 1957, for example) have sought to verify a rising rate of exploitation by reference to Marx's concept of unproductive labour (supervisory staff, bank employees etc.). If these workers are regarded as being paid out of surplus

value, rather than as constituting a cost of production which reduces surplus value, and if their relative importance in the labour force has been rising (which it has), then a rising rate of exploitation is consistent with a rising share of wages in national income. But to argue that the surplus value available to the capitalists for accumulation has declined because, given the growth of productivity of productive workers, there has been a growth in the proportion of unproductive workers, does not seem to add much to the simpler idea that the growth of productivity of all workers has been insufficiently fast relative to real wages.

The rising trend of real wages has raised the issue as to whether Marx's concept of the value of labour power, dependent on the time required to produce the 'necessaries' is still valid. The usual answer has been for Marxists to stress the 'moral and historical' element in the value of labour power as defined by Marx. Periods of strong demand for labour and the development of trade unions have allowed a widening of workers' 'needs', including the provision of more extensive state services. The difficulties that employers have found in cutting real wages, and governments in seriously eroding the welfare services despite the mass unemployment in the 1970s and 1980s, have added conviction to the idea that the current standard of living is, socially, *necessary* (Rowthorn 1980, ch. 7).

Marx argued that the trend towards a rising organic composition would allow the rate of exploitation to be increased, but would nevertheless lead to a falling *rate* of profit on total capital employed as outlays on constant capital would grow. Despite the fact that Marx regarded this Law of the Tendency of the Rate of Profit to Fall as the 'most important law of political economy' it played only rather a background role in the classic works of Marxism (Luxemburg 1913; Hilferding 1910). With the revival of interest in Marxist economics in the late 1960s it received prominence in the works of writers such as Mandel (1975). The main controversy has surrounded whether or not there is a fundamental tendency for the value of constant capital per worker to rise as the Law requires. Marx himself recognized that this was the outcome of a twosided process. The increased mass of constant capital per worker tended to drag the value of capital up. On the other hand the productivity growth which was part and parcel of the process tended to reduce the value of constant capital per worker. Whether the value of constant capital per worker rises or falls depends on whether productivity grows slower or faster than the increased mass of capital per worker. Marx himself gave no convincing reasons why productivity growth should be the slower of the two, and it has long been argued that there is no such reason (Robinson 1942; Sweezy 1942; van Parijs 1980). Attempts to argue that in some sense the rise in the mass of constant capital per worker is more fundamental and that there is a Law of the Tendency of the Rate of Profit to Fall even if it was manifested in an upward trend in the profit rate (Fine and Harris 1978) have not been found convincing. Marxists who have attempted to provide empirical evidence in support of the Law have typically confused the mass of constant capital with its value: the capital–output ratio, which is the price correlate of the value of capital per worker, has not shown an upward trend.

If this objection makes a falling profit trend contingent on the strength of productivity growth (an empirical matter), the second line of objection (originated by Okishio 1961) argues in fact that the techniques willingly introduced by capitalists will never, in and of themselves, result in a lower profit rate for the capitalist class. If can be shown that new techniques which raise the profit rate for the innovating capitalist will also imply, contrary to Marx's belief, a cost saving and thus a higher profit rate for the capitalist class. For the average profit rate to fall with the introduction of new techniques, therefore, there must have been, in addition, some increase in real wages. All this is not to say that the value of constant capital may not rise in some periods, and that it may not be associated with a falling profit rate (both were true of many countries in the early Seventies), but only that there must also be rising wages (as was also the case). It has been argued by Shaikh (1978) that oligopolists might not maximize the profit rate; but even if this were so it could not establish any necessity for the profit rate to fall.

Discussion of the Law of the Tendency of the Rate of Profit to Fall has emphasized the importance of the course of real wages for the development of capitalism. The two main schools of Marxist crisis theory have indeed placed real wages at the centre, but in very different ways. Underconsumptionist theorists (Luxemburg in the classic period, Sweezy amongst later writers) have argued that insufficient growth in real wages depresses the incentive to invest by restricting the market for consumer goods. As Tugan-Baranovsky (summarized by Sweezy 1942) pointed out with the help of Marx's reproduction schemes, it is not possible to prove the *necessity* of a crisis of underconsumptionism from a rising rate of surplus value. As Marx explained, whether or not surplus value was realized depends entirely on capitalists' spending decisions (on investment and consumption). The capitalists could realize a growing share of surplus value provided they were prepared to invest more and more in the capital goods sector (Dept I), even though this investment was destined just to produce more capital goods (Bukharin 1924). So crises of underconsumption, which would arise when capitalists failed to increase their investment in line with the potential surplus value, rely on the behavioural assumption that capitalists will actually not keep up their investment spending. The most influential postwar analysis along these lines, Baran and Sweezy's *Monopoly Capital* (1966, which acknowledges its theoretical debt to Steindl 1952), saw the growing monopolization of US capitalism enhancing the tendency for the share of surplus value to rise, while at the same time relaxing the pressure to invest.

It was something of an irony that just at the time that *Monopoly Capital* was written, Europe and Japan were enjoying a phenomenal boom. Many Marxist economists in these countries favoured an overaccumulation theory of crisis (Glyn and Sutcliffe 1972; Rowthorn 1980, chs 4–6; Itoh 1980). The strength of the boom eroded the reserve army of labour and caused tight labour markets, rising wages and thus falling profits, inflation and a recession (Armstrong et al. 1984). Also emphasized by these theories has been the role of stronger trade unions in pressing for higher state welfare spending and the difficulties that full employment brought for employers attempting to reorganize production to increase productivity (Bowles et al. 1983).

Why these difficulties should lead to a crisis, rather than simply slower growth, again depends on the central question of capitalists' investment behaviour. Precisely why and when a fall in profits leads to a precipitate decline in investment is notoriously difficult to model. Japanese Marxists (Itoh 1980) have made an important contribution by emphasizing the importance of the credit system in both prolonging a boom and initiating a collapse. Kalecki, who immortalized Marx's insight in the dictum 'workers spend what they get, capitalists get what they spend', wrote near the end of his life that the determination of investment 'remains the great *pièce de résistance* of economics' (1971, p. 165).

The recuperative role of crises in restoring the conditions for renewed accumulation has always been stressed by Marxists. It is more plausible in the case of crises due to overaccumulation (where the problem is rising wages) than for underconsumption crises (where wages have been rising too slowly). Indeed, Keynesian policies of demand expansion seem designed to meet the latter, and political difficulties have to be put forward as blocking such an obvious solution (Baran and Sweezy 1966). In crises of overaccumulation Keynesian policies are more likely to be used in reverse, in order to speed up the impact of unemployment in reducing labour's bargaining position over wages and productivity. Some French Marxists, known as the 'Regulation School' have recently emphasized the necessity for the whole pattern of institutions, state policies, technologies etc. to be reformed if a major structural crisis is to be overcome (Aglietta 1979; Boyer 1979; de Vroey 1984). Whether the microchip, decentralization of production, internationalization of production and capital markets, Japanesestyle industrial relations, more freedom for market forces and so forth provide a new 'way out' for capitalism in the 1990s is currently under intense discussion.

If this review of Marxist economics has concentrated on debates about, revisions to, and

extensions of Marx's own ideas it is to emphasize that the days of Stalinist orthodoxy and dogmatic repetition of the texts are gone. Marxist economics is again making a forceful and imaginative contribution to the analysis of contemporary society.

See Also

► Marx, Karl Heinrich (1818–1883)

Bibliography

Aglietta, M. 1979. *A theory of capitalist regulation*. London: New Left Books.

Armstrong, P., A. Glyn, and J. Harrison. 1984. *Capitalism since world war II*. London: Fontana.

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

Böhm-Bawerk, E. [1896]. 1948. *Karl Marx and the close of his system*, ed. P. Sweezy. New York: Kelly. (First published in German.)

Bortkiewicz, L. 1906. On the correction of Marx's fundamental theoretical construction in the third volume of Capital. In Böhm-Bawerk, (1948). (First published in German.)

Bowles, S. 1985. The production process in a competitive economy. *American Economic Review* 75(2): 16–36.

Bowles, S., D. Gordon, and T. Weisskopf. 1983. *Beyond the waste-land*. New York: Anchor Press.

Boyer, M. 1979. Wage formation in historical perspective: The French experience. *Cambridge Journal of Economics* 3(2): 99–118.

Braverman, H. 1972. *Labor and monopoly capital*. New York: Monthly Review Press.

Bukharin, N. 1924. In *Imperialism and the accumulation of capital*, ed. K. Tarbuck. London: Allen Lane, 1972. (First published in German.).

Dobb, M. 1937. *Political economy and capitalism*. London: Routledge & Kegan Paul.

Edwards, R. 1979. *Contested terrain*. London: Heinemann.

Fine, B., and L. Harris. 1978. *Rereading capital*. London: Macmillan.

Gillman, J. 1957. *The falling rate of profit*. London: Dobson.

Glyn, A., and B. Sutcliffe. 1972. *British capitalism, workers and the profit squeeze*. Harmondsworth: Penguin.

Gough, I. 1979. *The political economy of the welfare state*. London: Macmillan.

Hilferding, R. 1910. *Finance capital*. London: Routledge & Kegan Paul, 1981. (First published in German.).

Himmelweit, S., and S. Mohun. 1981. Real abstractions and anomalous assumptions. In *The value controversy*, ed. I. Steedman et al. London: Verso.

Itoh, M. 1980. *Value and crisis*. London: Pluto Press.

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economies*. Cambridge: Cambridge University Press.

King, J. 1982. Value and exploitation: Some recent debates. In *Classical and Marxian political economy*, ed. I. Bradley and J. Howard. London: Macmillan.

Luxemburg, R. 1913. *The accumulation of capital*. London: Routledge & Kegan Paul, 1951, (First published in German.).

Mandel, E. 1962. *Marxist economic theory*. London: Merlin.

Mandel, E. 1975. *Late capitalism*. London: New Left Books.

Marglin, S. 1984. *Growth, distribution and prices*. Cambridge, MA: Harvard University Press.

Medio, A. 1972. Profits and surplus value. In *A critique of economic theory*, ed. E. Hunt and J. Schwartz. Harmondsworth: Penguin.

Morishima, M. 1973. *Marx's economics*. Cambridge: Cambridge University Press.

O'Connor, J. 1973. *The fiscal crisis of the state*. New York: St Martin's Press.

Okishio, N. 1961. Technical change and the rate of profit. *Kobe University Economic Review* 7: 85–99.

van Parijs, P. 1980. The falling-rate of profit theory of crisis. *Review of Radical Political Economics* 12(1): 1–16.

Robinson, J. 1942. *An essay on Marxian economics*. London: Macmillan.

Rosdolsky, R. 1968. *The making of Marx's capital*. London: Pluto Press, 1977. (First published in German.).

Rowthorn, R. 1980. *Capitalism, conflict and inflation*. London: Lawrence & Wishart.

Rubin, I. 1928. *Essays on Marx's theory of value*. Detroit: Black & Red, 1972. (First published in Russian.).

Samuelson, P. 1971. Understanding the Marxian notion of exploitation. *Journal of Economic Literature* 9: 399–431.

Sen, A. 1978. On the labour theory of value. *Cambridge Journal of Economics* 2: 175–180.

Seton, F. 1957. The transformation problem. *Review of Economic Studies* 24: 149–160.

Shaikh, A. 1978. Political economy and capitalism. *Cambridge Journal of Economics* 2(2): 232–251.

Shaikh, A. 1981. The poverty of algebra. In *The value controversy*, ed. I. Steedman et al. London: Verso.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.

Steindl, J. 1952. *Maturity and stagnation in American capitalism*. Oxford: Blackwell.

Sweezy, P. 1942. *The theory of capitalist development*. New York: Monthly Review Press.

de Vroey, M. 1984. A regulation approach interpretation of the contemporary crisis. *Capital and Class* 23: 45–66.
Willman, P., and G. Winch. 1985. *Innovation and management control*. Cambridge: Cambridge University Press.
Woolf, E. 1979. The rate of surplus value, the organic composition of capital and the general rate of profit in the US economy 1947–67. *American Economic Review* 69(3): 329–341.

Mason, Edward Sagendorph (1899–1992)

Gustav F. Papanek

Keywords
Development economics; Firm, theory of; Industrial economics; Mason, E. S.; Monopolistic competition

JEL Classifications
B31

Edward Mason had a significant impact on the economics profession in four disparate areas: through his influence on the Harvard Economics Department; through his role in two separate subfields of the discipline – industrial and development economics; and by exemplifying the dual role of academic and practitioner.

When he came to Harvard in 1919 as a graduate student, he was not in the Harvard mould: he was from the public schools of Kansas and its University. Later, another newcomer and 'provincial', J. Kenneth Galbraith, characterized young Mason's presence by saying 'even when he was an Instructor, where Ed Mason sat there was the head of the table'. He remained a central figure in Harvard Economics for well over 50 years, his only absence a stint in Washington during the Second World War (1941–44). He was one of a handful of senior faculty who dominated the department during its glory days, when it produced about half of all economics Ph.D. degree holders in the United States and was responsible for a substantial fraction of the research produced in the country.

Mason's role at Harvard extended beyond economics. He served for 11 years (1947–58) as Dean of the School of Public Administration and for a short period was second-in-command of the university. While he was its Dean the School of Public Administration became the leading exponent of an emphasis on policy analysis, especially economic analysis, rather than administrative tools and institutions.

Mason taught many of the economists who expanded industrial economics from a preoccupation with regulation and monopoly to an analysis of markets and firms. The 'Masonic Lodge' of ex-students, together with its Grand Master, came to dominate the newly developed applied discipline of industrial organization. Mason stimulated the work by his proposition that the performance of a firm was largely explained by the structure of the market in which it operated. The controversy stimulated by this idea and by the concept of monopolistic competition to which he contributed, helped the subfield of industrial economics to flourish intellectually.

It was typical of Mason to come to research from policy concerns; he exemplified the practitioner–academician. Mason's advice was sought by governments and he was privy to their problems and the facts at their disposal. During the Second World War, he headed the economic staff of the OSS, probably the first US intelligence agency to gather economic intelligence systematically and to analyse possibilities for economic warfare. Later he was Deputy Assistant Secretary of State in charge of Economic Affairs and chief economic adviser to the US delegation at the 1947 Moscow Conference.

In the early 1950s, development economists began to apply many of the standard tools of economics to the special problems and institutions of the poor countries. Mason developed an interest in this subject in typical fashion by setting out to deal with a specific set of policy problems. After returning to Harvard at the end of the war, as Dean of Public Administration, he was asked to organize technical assistance to help Pakistan carry out the economic analysis crucial to rational government decision making. Out of that effort emerged another institution that rightfully could

be seen as created by Mason; the Development Advisory Service, later the Harvard Institute of International Development. A surprising proportion of those who consider themselves development economists had their initial field experience as an adviser, or as their local counter part, in one of the teams fielded by an organization that existed only because Mason managed to persuade Harvard to undertake the non-traditional university function of advising foreign governments.

Selected Works

1926. The doctrine of comparative cost. *Quarterly Journal of Economics* 41(November): 63–93.

1932. *The street railway in Massachusetts*. Cambridge, MA: Harvard University Press.

1937. Monopoly in law and economics. *Yale Law Journal* 47(1): 34–49. Repr. in *Public policy and the modern corporation: Selected readings*, ed. D. Grunewald and H.L. Bass. New York: Meredith, 1966.

1938. Price inflexibility. *Review of Economics and Statistics* 20(May): 53–64.

1939. Price and production policies of larger-scale enterprise. *American Economic Review* 29: 61–74. Repr. in American Economic Association, *Readings in industrial organization and public policy*, ed. R.B. Hefelbower and G.W. Stocking. Homewood: Richard D. Irwin, 1958.

1946. *Controlling world trade cartels and commodity agreements*. New York: McGraw-Hill.

1949a. The effectiveness of the federal anti-trust laws: A symposium. *American Economic Review* 39: 712–713. Repr. in *Monopoly power and economic performance: The problems of industrial concentration*, ed. E. Mansfield. New York: W.W. Norton, 1964. Also reprinted in *Problems of the modern economy*, ed. E.S. Phelps. New York: W.W. Norton, 1966.

1949b. The current status of the monopoly problem in the United States. *Harvard Law Review* 62: 1265–1285.

1952. Raw materials, rearmament, and economic development. *Quarterly Journal of Economics* 66: 327–341.

1956. Market power and business conduct: Some comments on the report of the Attorney General's Committee on Antitrust Policy. *American Economic Review, Papers and Proceedings* 46: 471–481.

1957. *Economic concentration and the monopoly problem*. Cambridge, MA: Harvard University Press.

1958. The apologetics of 'Managerialism'. *Journal of Business* 31(January): 1–11. Repr. in *The business system: Readings in ideas and concepts* (vols 1–3). *In commemoration of the fiftieth anniversary, Graduate School of Business, Columbia University, 1966*, ed. C.C. Walton and R.S. Eells. New York: Arkville Press, 1967.

1959., ed. *The corporation in modern society*. Cambridge. MA: Harvard University Press.

1960. The role of government in economic development. *American Economic Review, Papers and Proceedings* 50: 636–640. Repr. in *Studies in economic development*, ed. A.M. Okun and R.W. Richardson. New York: Holt, Rinehart & Winston, 1961.

1962. Some aspects of the strategy of development planning: Centralization vs. decentralization. In *Organizations, planning and programming for economic development*, vol. 8 of Science, Technology and Development; US papers prepared for the U.N. Conference on the Application of Science and Technology for the Benefit of the Less Developed Area, Washington, DC: Government Printing Office.

1963. Interests, ideologies and the problem of stability and growth. *American Economic Review* 53: 1–18.

1964. *Foreign aid and foreign policy*. New York: Harper and Row for the Council on Foreign Relations.

1966. *Economic development in India and Pakistan*. Cambridge, MA: Center for International Affairs, Harvard University.

1967. Monopolistic competition and the growth process in less developed countries: Chamberlin and the Schumpeterian dimension. In *Monopolistic competition theory: Studies in impact. Essays in honor of Edward H. Chamberlin*, ed. R.E. Kuenne. New York: Wiley.

1971. Controlling industry. In D.V. Brown et al., *The economics of the recovery program*. New York: Da Capo.
1973. (With R.E. Asher.) *The World Bank since Bretton Woods*. Washington, DC: Brookings Institution.
1980. (With others.) *The economic and social modernization of the Republic of Korea*. Cambridge, MA: Harvard University Press

Massé, Pierre (1898–1987)

Marcel P. Boiteux

Keywords
Bellman's Optimum Principle; Centre and periphery; Dynamic programming; Electricity markets; Forecasting; Linear programming; Massé, P.; Optimum control; Planning

JEL Classifications
B31

Massé was born on 13 January 1898, the same day that Emile Zola in his 'J'accuse' revealed the truth about the Dreyfus Affair, which arouses so much passion in French political circles to this day. His family was quick to take sides – in defence of the innocent – and from them Massé quite probably inherited his deep humanism, in which realistic thought was allied with optimism in action.

In 1916, he passed the competitive entrance examinations to both the Ecole Normale Supérieure, in science, and the Ecole Polytechnique. He was a Second Lieutenant in the Artillery from 1917 to 1918, and then opted for the Polytechnique and a career as an engineer. This choice of career foreshadowed a life in which, in a happy marriage, thought and action were to mingle unceasingly. A further spell of training at the Ecole des Ponts et Chaussées, and the start of his career as a government servant, channelled him towards major civil engineering works, and then, quite soon, towards the business world and hydroelectrical improvement works.

This was a decisive turning-point for both the man of action, the builder, and for the thinking economist. Obliged to deal with the management problems raised by the water stocks accumulated in reservoirs, and also with the need to turn them to account, Massé identified the key role of reserves as the means of regulating systems in order to cope with random factors. In his first work, *Les réserves et la régulation de l'avenir*, published in 1946, whose findings had been published two years previously in a paper submitted to the Société Statistique de Paris, Pierre Massé can be seen to be a forerunner of dynamic programming and of the theory of optimum control. In particular, he set forth in this paper two rules for the optimal management of random processes: *(a)* reservoirs should be managed so as to equalize the marginal utility of the water releases and the marginal expected value of the water held in stock; and *(b)* in order to calculate that expected value a strategy for the future should be defined, that is, a sequence of conditional decisions combining at any time the impact of past decisions, the actual outcome of the random processes, and the perception of what future natural conditions will probably be. Kenneth Arrow was later to note, in 1956, that this was the earliest formulation of Richard Bellman's Optimum Principle.

Massé's work was deeply marked by the recognition that in a random world – and the more so with an uncertain future – one could not confine oneself to just a single forecast, and by the need to adopt strategies and regulate stocks. This was to be borne out 20 years later, when he was in charge of the French Commissariat au Plan (Planning Commission). The consistency of the forecasts carried out as part of the National Accounts exercise certainly went some way towards making the plan a 'reducer of uncertainties'. Moreover, under his guidance this achievement was crowned by a forecasting approach in which the seeking of a consensus on the type of development that was desirable was combined with the concern to identify 'factors with potential for the future', and by the devising of 'warning lights',as instruments for marking the

future course that were capable of setting corrective actions in motion.

He was Directeur de l'Equipement at Electricité de France in 1946 for the start of the Plan Monnet and became its Directeur Général Adjoint two-and-a-half years later, a post he held until 1959. In those 12 years he developed, and then applied, linear programming techniques for determining the overall volume of electricity generating plant, and furnished justification for using a national discount rate for setting off present and future income and expenditure against each other. He tirelessly argued with the government in favour of using these clear and rigorous tools, already finding support on the Commissariat Général du Plan. In 1957, he published *Le choix des investissements,* a work which was to become authoritative both in France and abroad.

In February 1959, General de Gaulle appointed Massé to head the Commissariat Général du Plan. He took up his duties backed by the sound experience of a microeconomist who was thoroughly at ease with the idea of maximizing the benefit to the community in managing a public service, and who was attached to the pricing system and to its role in providing guidance and regulation. He sought to make the Plan – which had been largely governed by the concern for consistency and accordingly gave pride of place to analysing interlocking strengths and weaknesses – a structure better directed towards achieving competitiveness, both domestically and on foreign markets. His aim was not only to produce more, but also to produce better quality, with consciousness of costs.

With these goals, he strove to lighten the Plan's structure and make it interlocking with, and not a substitute for, the market. Without losing the valuable contribution of a generalized market survey, backed up with the use of an input–output matrix, he endeavoured to better pinpoint future price and income trends: in this way, programming by volume was to be backed up by an early attempt at programming by value. While the market could show what present prices were, it said practically nothing about future prices, since forward markets covered only narrow sectors and near time-horizons. By the light it shed on the future, the Plan was seen to be an indispensable adjunct to a smoothly working market economy. The 'Centre' had the task of successfully conveying to the 'Periphery' the right price system, and on the basis of this information, the Periphery was able to return to the Centre information on what the intentions were of the decentralized economic agents concerning volumes of goods to be consumed or invested in, and the volumes of factors to be mobilized. In this way, consultation was established between the Centre and the Periphery, converging, after a few successive iterations, towards a dynamic equilibrium. Pierre Massé had already analysed this converging dialogue between the Centre and the Periphery as early as 1952, in 'Pratique et philosophie de l'investissement'. The Commissariat au Plan in fact organized consultations among the major socio-occupational categories; experts could intervene to put figures to the impacts from selecting the options adopted, while the representatives of the state were there to ensure observance of the major policy guidelines defined by the government in agreement with Parliament. Such at least was the theory. Certain departures from it were unavoidable in practice.

However, 'at the same time as it was an act of faith', the Plan continued to be 'an affirmation of the will'. Concerned as he was for a 'less incomplete view of man' to be taken into consideration, Massé succeeded in convincing the most influential circles in his country that a better balance should be struck between private and collective consumption. Thus, a feature of the early 1960s was a new concern for developing communal infrastructures. At the same time, while investigating various development scenarios, he concluded that it was necessary to raise the discount rate (of profitability) – an indicator of the scarcity of capital for government investors – so that it actually corresponded to the marginal efficiency of capital. He also concerned himself with disseminating the practice of constant-price calculations, so that, while changes in relative prices were not ignored, the profitability of infrastructure projects was not made attributable to inflated profits.

Having stressed future values, Massé necessarily broadened the scope of studies to cover price and income trends, and unavoidably brought discussion round to the knotty point of social

tensions. To clarify and persuade, he worked on surplus accounts, establishing a rigorous relationship between the overall productivity gains made from one year to the next, and the sum of benefits available for distribution to customers, suppliers, workers and investors. From this attempt there at least remains a learning approach which the Centre d'Etude des Revenus et des Coûts, set up in 1966 at his instigation, has been engaged in disseminating and extending.

After helping start the Fifth Plan, Massé returned to Electricité de France, of which he was chairman for three years, and secured from the political authorities a sounder channelling of the necessary efforts for investment in the generation of electricity by nuclear power. He thereupon resumed acquaintanceship with business economics, though he did not forsake reflections upon the problems of the national economy, which he was never to abandon thereafter.

In 1977, Massé was elected a member of the Institut de France, which for almost 200 years has gathered together the most eminent French personalities in the humanities, science, history, philosophy and art. He pursued research for the remainder of his life. His body of work attests to a lifelong endeavour to reconcile macro and micro-economists and to ensure the cross-fertilization of their ideas for the social good.

Selected Works

1946. *Les résberves et la régulation de l'avenir dans la vie économique.* Paris: Hermann.
1952. Pratique et philosophie de l'investissement. *Economie appliquée* 5: 625–658.
1953. Les investissements électriques. *Revue de statistique appliquée* 1(3–4): 119–129.
1957. (With R. Gibrat.) Application of linear programming to investments in the electric power industry. *Management Science* 3(2): 149–66.
1959. Prévision et prospective. *Cahiers de prospective*. Paris.
1964. *Le choix des investissements*, 2nd ed. Paris: Dunod.
1965. *Le plan ou l'anti-hasard*. Paris: Gallimard.
1969. (With P. Bernard.) *Les dividendes du progrès*. Paris: Le Seuil.
1973. *La crise du développement.* Paris: Gallimard.
1976. Prédation et création. *Etudes*. Paris.
1982. Le chiffré et le vécu, 1960–1980. *Revue des sciences morales et politiques*. Paris.
1984. *Aléas et Progrès, entre Candide et Cassandre*. Paris: Economica.

Massie, Joseph (Died 1794)

Murray Milgate

Joseph Massie is one of those figures of mid-18th-century economics about whose life little of real substance is known, but who has nevertheless managed to attract from quite diverse authors a degree of recognition that is not enjoyed by many much less mysterious figures. In *Theories of Surplus Value* Marx praises Massie for the subtlety of his discussion of the general rate of profit; the original edition of this *Dictionary* credits him with a decisive criticism of Locke's theory of the rate of interest; William Cunningham admired Massie's belief that 'only an exhaustive examination of the phenomena of industrial and commercial life' could yield an appropriate understanding of the principles which governed economic activity (1891, p. 81); and historians of economic thought are thankful for his extraordinary activities as a book-collector that yielded a catalogue to a personal library (containing by 1764 more than two thousand items) of economic literature prior to 1750.

From the point of view of economic theory, there would seem to be little doubt that Massie's most important contribution is to be found in his *Essay on the Governing Causes of the Natural Rate of Interest* (1750), which was reprinted with an introduction by Jacob Hollander in 1912. However, quite what it is that makes this short treatise so important is a matter of debate. Some argue that Massie's rejection of the general argument of Locke, that the money rate of interest regulates

the profits of trade, is its key feature. Certainly, the emergence of this line of criticism played a part in shifting the focus of the study of the rate of profit from the sphere of money to the sphere of trade in general – a transition which was to be important in the construction of classical economic theory by Adam Smith and others. On this point, some (like Marx) have been argued that David Hume, who published a more widely canvassed expression of the same ideas after Massie, 'borrowed' them from that author – thereby crediting Massie with a direct, if unacknowledged, influence on the development of economic theory.

However, there is another feature of Massie's discussion of the governing causes of the natural rate of interest which has caught the attention of others. This is his idea that in a competitive economy the *natural* level of the rate of profit is that to which the profits of different trades tend to conform – it was, as Smith was to put it twenty-five years later, the centre of gravitation of the system. Moreover, according to Massie, when the returns to investment in different lines of production are at this natural level, commodities sell for what he calls their *natural price*. The significance of this is, of course, not that Massie used the notion of natural price, but rather that he *defined* natural prices in terms of a general rate of profit.

Massie, however, does not seem to have appreciated that these radically new conceptualizations could provide the basis upon which to build economic arguments of a more systematic, or scientific, character than had hitherto been the case – as Smith, for example, was to do to so great an effect twenty-six years later in the *Wealth of Nations*. Indeed, Massie actually wrote on the question of systematizing economic study, and here his arguments seem to have little in common with the grand theorists. As Cunningham was the first to observe, Massie's *Representation concerning the knowledge of Commerce* (1760a) advocated instead a detailed historical investigation of the actual conditions of industry – so detailed, in fact, that in his account of each branch of industry there was to be sixteen sub-divisions into which each was to be broken. Even for Cunningham, this was pushing the historical method a little too far.

Selected Works

1750. *An essay on the governing causes of the natural rate of interest*. London. Reprinted and edited by J. Hollander. Baltimore: Johns Hopkins Press, 1912.
1756. *Calculations of taxes for a family of each Rank, degree or class: for one year*. London.
1758. *A plan for the establishment of charity-houses for exposed or deserted women and girls, and for penitent prostitutes, … etc*. London.
1760a. *A representation concerning the knowledge of commerce as a national concern*. London.
1760b. *Observations relating to the coin of Great Britain, … & etc*. London.
n.d. *Alphabetical index of the names of authors of commercial books and pamphelets*. British Museum, Landsdowne MS. No. 1049.

Bibliography

Cunningham, W. 1891. Economic doctrine in the eighteenth century. *Economic Journal* 1: 73–95.

Matching

Giuseppe Moscarini

Abstract

Matching (or job-matching) is the process whereby a firm and a worker meet, learn whether their characteristics combine productively and, in light of this information, sequentially contract a wage and decide whether to separate or to continue production. This hypothesis implies that wages rise and the risk of separation declines with seniority, wage changes are unpredictable and have declining variability, and valuable specific human capital is accumulated in the form of knowledge about the quality of the match.

These and other observable implications have found strong support in available empirical evidence, and make job-matching a central theory of worker turnover.

Keywords

Bellman equation; Job-matching hypothesis; Labour-market contracts; Matching; Matching markets; Returns to tenure; Roy model; Selection; Wage distribution; Worker turnover

JEL Classifications

J410

Matching (or job-matching) is the process whereby a firm and a worker meet, learn whether their characteristics combine productively and, in light of this information, sequentially contract a wage and decide whether to separate or to continue production.

In many respects, a job is like a marriage. Two parties (a firm and a worker) engage in a long-run relationship, whose success depends on a myriad of factors, all quite difficult to describe. Only the actual outcome of the match can reveal the underlying 'fit'. If the match works, it continues; otherwise it is scrapped and the partners try their luck elsewhere.

Jovanovic (1979a) formalizes the job-matching hypothesis in a dynamic, rational-expectations context. This hypothesis hinges on two pivotal ideas: learning and selection. The emphasis on selection follows the tradition of equilibrium sorting in labour markets going back to the static Roy model (1951). Now, dynamics and imperfect information take centre stage. A job is viewed as an 'inspection' as well as an 'experience' good. The worker and the firm have to 'taste' the match to decide its value, just like two people first date (to 'inspect' the match) then possibly get married (to 'experience' the match), with varying degrees of success. Unlike in marriage markets, utility is typically transferable through the wage. The fit between firm and worker characteristics is modelled as a match-specific productivity component, a parameter of the output process, summarizing how well the innumerable relevant characteristics of the worker and of the task actually dovetail. Random noise in production creates a signal extraction problem. The firm and/or the worker continuously observe the output performance of the match, incorporate this information in wages, and reassess it against alternative opportunities offered by the market.

A Job-Matching Model

Output y_t is produced at time $t = 1, 2\ldots$ by a firm and a worker with a 1:1 Leontief technology:

$$y_t = \theta + \varepsilon_t.$$

There is no hours or effort choice. θ is average productivity or 'match quality', drawn by nature, unobserved by firm and worker, at the beginning of the match from $\theta : N(m_{-1}, 1/h_{-1})$, which are also parties' prior beliefs. $\varepsilon_t : N(0, 1/p_\varepsilon)$ is white noise, i.i.d. and independent of θ. Therefore, risk-neutral firm and worker are interested in the permanent component θ. Following the bulk of the literature, assume that firm and worker are symmetrically informed. This is not a crucial assumption: all that matters is that *some* learning drives match selection.

Upon matching at time 0, parties inspect the match and observe a signal

$$x = \theta + \eta$$

where $\eta : N(0, 1/p_\eta)$ independent of θ. By Bayes' rule, $\theta|x : N(m_0, 1/h_0)$, where $h_0 = h_{-1} + p_\eta$ and $h_0 m_0 = m_{-1}\, h_{-1} + x p_\eta$. If the match begins and output is produced at $t = 1, 2\ldots$, posterior beliefs about match quality conditional on the worker's track record are recursively updated as follows:

$$\begin{aligned} &\theta|x, y_1, y_2, \quad \ldots y_t : N(m_t, \ \ 1/h_t) \text{ where} \\ &h_t = h_{-1} + p_\eta + t p_\varepsilon \ \ h_t m_t \\ &\quad = h_{t-1} m_{t-1} + p_\varepsilon y_t. \end{aligned}$$

That is, m_t and h_t are the mean and precision of the normal posterior distribution of θ, conditional on all information available to date t. After solving

backward, m_t is an average of the prior expectation m_{-1}, the initial signal x and the history of output $\sum_{s=1}^{t} y_s$, weighted by their respective precisions. Given the model's parameters, history and beliefs are summarized by expected productivity m_t and by tenure t, which jointly measure the specific human capital accumulated in the relationship.

With no uncertainty and perfect information ($h_{-1} = \infty$ and/or $p_\eta = \infty$), workers and firms would immediately discard unpromising matches and keep drawing better and better outcomes. With imperfect information, equilibrium behaviour is 'sequential' and non-trivial. Equilibrium cannot be perfectly competitive, due to the specificity of match quality and consequently of human capital. Nonetheless, with free entry, no mobility and no capital costs, there is a contracting equilibrium where the wage offered by the firm to the worker equals the worker's expected (marginal) productivity m_t, and firms break even. The worker captures the entire option value of learning. By Bayes's rule, the distribution of the future wage m_{t+1}, unconditional on unknown match quality θ but conditional on current beliefs $\{m_t, t\}$, is normal with

$$\begin{aligned} \mathrm{E}[m_{t+1}|m_t,t] &= m_t \text{ and } Var[m_{t+1}|m_t,t] \\ &= \frac{p_\varepsilon}{\left[h_{-1}+p_\eta+(t+1)p_\varepsilon\right]\left(h_{-1}+p_\eta+tp_\varepsilon\right)} \end{aligned} \tag{1}$$

The worker's value of employment solves the Bellman equation

$$\begin{aligned} V(m_t,t) = \max\langle &\beta \mathrm{E}[V(\tilde{m}_0,0)], \\ &m_t + \beta \mathrm{E}[V(m_{t+1},t+1)|m_t,t\rangle \end{aligned} \tag{2}$$

for some discount factor $\beta \in [0,1]$. At each point in time, including $t = 0$ right after observing the initial signal x and before starting production, the worker decides whether to quit this match at once and to inspect another one next period (expected value $E[V(\tilde{m}_0,0)]$, independent of $\{m_t, t\}$ because θ is match-specific) or to accept the wage m_t, produce, observe the output realization y_t, update beliefs to $\{m_{t+1}, t+1\}$, and decide again.

The worker's employment value $V(m, t)$ is increasing in expected match quality m and decreasing in tenure t. The first effect is obvious. Formally, an increase in m_t raises the right-hand side of (2) directly and, by (1), the normal distribution of future wages in a first-order stochastic dominance sense. Standard dynamic programming arguments establish monotonicity of V. To see why the value V is also decreasing in tenure t, consider the following thought experiment. Before deciding whether to quit or to produce y_{t+1}, the worker is provided with a free signal υ which has the same distribution as y_{t+1}, and is then informative about match quality. After observing this signal, the worker cannot do worse, because he or she can always ignore it. So, before observing υ, she must value this additional information:

$$\begin{aligned} \mathrm{E}_\upsilon[\mathrm{E}[V(m_{t+1},t+1)|\upsilon,m_t,t]] &\geq V(m_t,t) \\ = V(\mathrm{E}_\upsilon[\mathrm{E}[m_{t+1}|\,m_t,t,\upsilon]],t) & \end{aligned}$$

where the equality follows from $\upsilon : y_{t+1}$ and then $\mathrm{E}_\upsilon[\mathrm{E}[m_{t+1}|\mathrm{m}_t, t, \upsilon]] = \mathrm{E}[m_{t+1}|m_t,t] = m_t$. The inequality implies that V is convex in m. Since tenure t reduces the variance of m from (1), it follows from Jensen's inequality that $V(m,t)$ declines in t for given m. Intuitively, a match of equally expected but more uncertain productivity is more valuable: there is some chance it will turn out to be great, otherwise it can always be scrapped.

Testable Implications and Empirical Evidence

The key implications of the model derive from selection and learning, and those implications that are testable have indeed found strong empirical support.

Selection

Given the properties of V, the worker quits as soon as the wage falls short of a reservation wage, which is increasing in tenure. As the option value of learning is consumed, a given expected match quality is no longer sufficient to support the

match. Reservation wages are not directly observable, but the resulting selection does have indirect, testable implications. Only promising matches survive, so the average m_t (wage) in continuing jobs increases (cross-sectionally) with tenure t. Indeed, seniority has modest but consistently positive wage returns (Altonji and Williams 2005). As better matches are less likely to end, the hazard rate of separation, after an initial 'discovery' phase, declines with tenure, a very robust stylized fact (Farber 1994). Finally, censoring bad matches skews the distribution of wage residuals, conditional on observable worker and firm exogenous characteristics: a symmetric and thin-tailed Gaussian distribution of output turns into a distribution of 'unexplained' wages with a thick Pareto upper tail (Moscarini 2005), as in a typical empirical wage distribution.

Learning

From (1), unconditional on the unobserved quality of the match, the wage m_t is a martingale, with variance of innovations declining with tenure t. Beliefs updated in a Bayesian fashion cannot be expected to drift in any direction, for the same reason that asset prices are a random walk in efficient financial markets. Thus, unconditionally on tenure, within-job wage changes are uncorrelated and, as uncertainty about match quality is resolved, have declining variance (Mortensen 1988). Wage growth slows down over the course of a career. Indeed, the search for serial correlation in wage changes has been inconclusive, but the slowdown of wage growth is prevalent (Topel and Ward 1992). The wages of a cohort of workers 'fan out', as some workers are luckier than others and find earlier a good match that pays a high wage, and as commonly observed empirically. When a match separates due to an exogenous layoff (not modelled here, but easy to accommodate), the worker loses the entire match-specific human capital, so she suffers a persistent wage loss. This fully agrees with the available evidence (Jacobson et al. 1993). More problematic is the prediction (Mortensen 1988) that, as $V(m,t)$ falls with t, separation rises with tenure given the wage: empirical evidence (Topel and Ward 1992) suggests the opposite.

Alternative Hypotheses About Worker Turnover

In light of its intuitive appeal and empirical success, job-matching has become the benchmark model of worker turnover. It has in part inspired the canonical search- and-matching model of the labour market (Mortensen and Pissarides 1994), where *ex post* idiosyncratic uncertainty drives job flows while search frictions account for involuntary unemployment. But, despite its vast influence, the job-matching approach still faces alternative and competing views of worker turnover, which provide conceptually quite different explanations for the same set of stylized facts. The starker contrast is with pure search models, which may dispose of heterogeneity altogether. In the search literature, wage dispersion and dynamics originate from firms' power of monopsony and commitment to contracts, due to purely strategic considerations. Retention concerns and counter-offers (Burdett and Mortensen 1998; Burdett and Coles 2003) explain returns to seniority, declining separations rate and so forth. Closer to the job-matching approach is a class of models that retain heterogeneity and selection, but allow for the quality of the job to change physically over time, while in the job-matching model everything is predetermined, and parties only have to learn their fate. Notable examples are firm-specific training (Jovanovic 1979b) and learning-by-doing, as well as stochastic match-specific productivity shocks (Mortensen and Pissarides 1994). In these models, general properties of Bayesian learning, like the declining variance of innovations, must be assumed as ad hoc properties of the productivity process. Nonetheless, this lack of identification poses a formidable challenge, and motivates an ongoing research effort.

See Also

- ▶ Assortative Matching
- ▶ Bandit Problems
- ▶ Learning in Macroeconomics
- ▶ Matching and Market Design
- ▶ Roy Model

- ▶ Search Theory
- ▶ Selection Bias and Self-Selection
- ▶ Sequential Analysis

Bibliography

Altonji, J., and N. Williams. 2005. Do wages rise with job seniority? A reassessment. *Industrial and Labor Relations Review* 58: 370–397.
Burdett, K., and M. Coles. 2003. Equilibrium wage-tenure contracts. *Econometrica* 71: 1377–1404.
Burdett, K., and D. Mortensen. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39: 257–273.
Farber, H. 1994. The analysis of interfirm worker mobility. *Journal of Labor Economics* 12: 554–593.
Jacobson, L., R. Lalonde, and D. Sullivan. 1993. Earnings losses of displaced workers. *American Economic Review* 83: 685–709.
Jovanovic, B. 1979a. Job matching and the theory of turnover. *Journal of Political Economy* 87: 972–990.
Jovanovic, B. 1979b. Firm-specific capital and turnover. *Journal of Political Economy* 87: 1246–1260.
Mortensen, D. 1988. Wages, separations, and job tenure: On-the-job specific training or matching? *Journal of Labor Economics* 6: 445–471.
Mortensen, D., and C. Pissarides. 1994. Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61: 397–415.
Moscarini, G. 2005. Job matching and the wage distribution. *Econometrica* 73: 481–516.
Roy, A. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3: 135–146.
Topel, R., and M. Ward. 1992. Job mobility and the careers of young men. *Quarterly Journal of Economics* 107: 439–479.

Matching and Market Design

Muriel Niederle, Alvin E. Roth and Tayfun Sönmez

Abstract

Matching is the part of economics concerned with who transacts with whom, and how. Models of matching, starting with the Gale–Shapley deferred acceptance algorithm, have been particularly useful in studying labour markets and in helping design clearing houses to fix market failures. Studying how markets fail also gives us insight into how marketplaces work well. They need to provide a thick, uncongested market in which it is safe to participate. Clearing houses that do this have been designed for many entry-level professional labour markets, for the assignment of children to public schools, and for exchange of live-donor kidneys for transplantation.

Keywords

Centralized matching; Clearing houses; Congestion; Kidney exchange; Labour markets; Market design; Marriage markets; Matching; Medical labour markets; National resident matching program; School choice; One-sided and two-sided markets; Priority algorithms; Revelation of preferences; Strategy-proof allocation mechanisms; Two-sided markets

JEL Classifications

C78; L1

'Matching' is the part of economics that focuses on the question of who gets what, particularly when the scarce goods to be allocated are heterogeneous and indivisible; for example, who works at which job, which students go to which school, who receives which transplantable organ, and so on. Studying how particular matching markets succeed at creating efficient matches, or fail to do so, has yielded insights into how markets in general work well or badly.

Because market failures have sometimes been successfully fixed by devising new rules for both centralized and decentralized market organization, matching has been a major focus of the emerging field of market design. Some designs by economists have included labour market clearing houses for doctors and other health-care workers in the United States, both for their first jobs and as they enter specialties. Clearing houses have also been implemented in less traditional markets, which cannot adjust prices or wages to help clear the market, such as the matching of students to schools in New York City and in Boston. And new clearing houses are being

implemented for the organization of live-donor kidney exchanges among patients in need of a kidney transplant who have willing donors with whom they are incompatible.

In the next section we review some studies of matching, including some market failures that have been addressed either by introducing appropriate rules to a decentralized market (as in admissions to graduate programmes in American universities), or by introducing a centralized clearing house (as in the markets for new doctors in the United States, Canada, and Britain). The subsequent two sections consider the simple theory behind some clearinghouse designs. Then we return to some of the successful market design applications, which build on the theoretical models, but handle practical problems that are sometimes not yet fully understood in theory.

We focus on three kinds of market failure that sometimes impede efficient matching.

1. Failure to provide *thickness;* that is, to bring together enough buyers and sellers (or firms and workers, schools and students, and so forth) to transact with each other.
2. Failure to overcome the *congestion* that thickness can bring, that is, that can result when lots of buyers and sellers are trying to transact. That is, failure to provide enough time, or failure to make transactions fast enough so that market participants can consider enough alternative possible transactions to arrive at satisfactory ones.
3. Failure to make it *safe* for market participants to reliably reveal or otherwise act on their information.

Some Market Failures and Their Consequences

Unravelling, Congestion and Centralized Clearing Houses

A variety of professional labour markets have suffered from the *unravelling* of appointment dates: from year to year, appointments were made earlier and earlier in advance of actual employment. Markets that had once been thick, with many employers and applicants on the market at the same time, became thin, as potential employees faced early offers, dispersed in time, to which they had to respond before they could learn what other offers might be forthcoming. That is, applicants often received 'exploding' offers that had to be accepted or rejected without waiting to see whether a more desirable offer might be forthcoming. An applicant who accepts such an offer, in the case that acceptances are binding, will never learn of the more desirable offers that might have become available, but if the offer is reasonably desirable rejecting it might be very risky. And, when applicants are quickly accepting offers in this way, employers, when they make offers, have to start taking into account whether the offer is likely to be accepted, since by the time an offer is rejected other desirable applicants may have already accepted offers elsewhere. This often makes unravelling a dynamic process, with offers being made earlier and shorter in duration from year to year. This kind of unravelling has been described in detail in markets for lawyers (Avery et al. 2001), gastroenterologists (Niederle et al. 2006) and many others (see Roth and Xing 1994). A clear example is the market for new doctors (Roth 1984).

The first job for almost all new doctors in the United States and Canada is as an intern or resident at a hospital. In the early 1900s, medical graduates were hired for such jobs near the end of their fourth year of medical school, just before graduation. By the 1930s, hiring was largely completed half a year before graduation, and by the 1940s it had moved to sometimes as much as two years before graduation. That is, in the early 1940s, students were being hired long before they would begin work, at dispersed times, and without much opportunity to consider alternatives, and long before they had sufficient experience to reveal either to employers or to themselves what kinds of medicine they would most prefer and be best able to practise. There was widespread recognition among the participants that the market was often failing to create the most productive matches of doctors to hospitals, both because there was too little opportunity to consider alternatives and because the matching was being done

before important information about students became available.

One way in which many markets tried to address this failure was by attempting to establish rules concerning when offers could be made. In the market for new American doctors, the most concerted attempt at this kind of solution began in 1945 with the help of the medical schools, which agreed not to release any information to hospitals about students until a specified date.

However, the market experienced *congestion* in that hospitals found that they did not always have enough time to make all the offers they would like if their first offer was declined. Over the next few years students were called upon to make increasingly prompt decisions whether to accept offers. In 1945 offers were supposed to remain open for 10 days. By 1949 a deadline of 12 hours had been rejected as too long. Hospitals were finding that, if an offer was rejected after even a brief period of consideration, it was often too late for them to reach their next most preferred candidates before they had accepted other offers. Even when there was a long deadline much of this action was compressed into the last moments, because a student who had been offered a position that wasn't his first choice would be inclined to wait as long as possible before accepting, in the hope of eventually being offered a preferable position. So hospitals felt compelled to pressure students to reply immediately, and offers conveyed by telegram were frequently followed by phone calls requesting an immediate reply.

Congestion can be a problem in any market in which transactions take some time, but it is especially visible in entry-level professional labour markets in which many workers and jobs become available at the same time (for example, after graduation from university, medical school, law school, and so on).

In the face of congestion, many markets unravel, as employers try to gain more time to make offers by starting to do so earlier (Roth and Xing 1994). But the market for new doctors found a solution in the form of a centralized clearing house. Starting in the early 1950s, the various medical groups organized a centralized clearing house, which remains in use today, having undergone some changes over the years. Nowadays, a medical student applies to hospitals and goes on interviews in the winter of the final year of medical school, and then in February submits an ordered preference list of positions to the centralized clearing house, the National Resident Matching Program (NRMP). At the same time, the residency programmes (the employers) submit an ordered preference list of candidates. Once all the preference lists are collected, the clearing house uses an algorithm to produce a match, and residency programmes and applicants are informed to whom they have been matched. Although this clearing house began as an entirely voluntary one, it has been so successful that today it is virtually the only way that most residencies are filled. As we will see below, that success depends critically on the matching algorithm.

The NRMP, and clearing houses like it, also make very clear the kinds of issues involved for a marketplace to make it safe for participants to reveal their information. In a clearing house in which you are asked to state your preferences, the question is simply: is it a good idea to state your true preferences, or would you do better otherwise? For the NRMP we'll see that stating true preferences is indeed both safe and sensible. We'll also discuss clearing houses that failed this test, like the one for placing students into schools in Boston that consequently failed to accomplish their objectives in other ways also, and were redesigned.

Before presenting some formal models that will allow us to start to explain which matching algorithms and clearing houses have been successful and which have failed, it will be helpful to think about several different kinds of matching markets.

Two-Sided and One-Sided Matching Markets

Labour markets, like the market for new doctors, are usually modelled as two-sided markets, in which agents on one side of the market (workers) need to be matched with agents on the other side (employers), and each agent has preferences over possible matches. We'll see below that this two-sided structure allows strong

conclusions to be drawn about the properties of matchings and matching mechanisms.

In many markets this two-sided structure is absent. One way this occurs is when any participant in the market can be matched with any other. For example, if a group of people want to form pairs to be roommates or bridge partners, any one of them can in principle be matched with any other, although not all matchings would be efficient. We encounter markets of this kind when we speak of kidney exchange.

Another way in which markets can be one-sided is if the agents in the market need to be matched to objects, for example when people need to be assigned rooms in a dormitory, or places in a public school that doesn't itself have preferences or take strategic actions (unlike in a two-sided matching market). That is, such a market matches people to places, but only one side, the people, are active participants in the market. Some markets can also be hybrids, with both two and one-sided properties (as when schools aren't strategic players, but still have priorities over students).

Below we consider some static models of two and one-sided matching that have proved useful in the design of clearing houses, and in understanding what they do. In the section on design, we'll also speak about some decentralized design solutions to various market failures, such as unravelling. While there has been some good initial progress on formal models of decentralized markets, and dynamic models in which phenomena like unravelling can play out over time (see for example, Li and Rosen 1998), these areas are still in need of development, and have not yet received the theoretical attention commensurate with their importance in the study of markets generally (though see Niederle and Yariv 2007).

Formal Models of Matching

Two-Sided Matching Models

The workhorse models of two-sided matching come in several varieties. The simplest, presented in detail below, is the 'marriage model' in which each firm seeks to hire only a single worker, and wages and other kinds of price adjustment are represented simply in the preferences that workers and firms have for each other (for example, in these models, wages are part of the job description that determines preferences). However the kinds of results we present here generalize to models in which wage and price formation is explicitly included, some pointers to such models are included in the references.

The marriage model consists of two disjoint sets of agents, men $= \{m_1,\ldots, m_n\}$ and women $= \{w_1, \ldots, w_p\}$, each of whom has complete and transitive preferences over the agents on the other side (and the possibility of being unmatched, which we model as being 'matched to yourself'). Preferences can be represented as rank order lists, for example, if man m_i's first choice is w_3, his second choice w_2 [$w_3 > m_i\ w_2$] and so on, until at some point he prefers to remain unmatched, that is $P(m_i) = w_3, w_2, \ldots m_i \ldots$. If agent k (on either side of the market) prefers to remain single rather than be matched to agent j, that is, if $k > k\ j$, then j is said to be *unacceptable* to k. If an agent is not indifferent between any two acceptable mates, or between being matched and unmatched, we'll say he/she has *strict* preferences.

An outcome of the game is a *matching:* μ: M ∪ W → M ∪ W such that $w = \mu(m)$ iff $\mu(w) = m$, and for all m and w either $\mu(w)$ is in M or $\mu(w) = w$, and either $\mu(m)$ is in W or $\mu(m) = m$. That is, a matching matches agents on one side to agents on the other side, or to themselves, and if w is matched to m, then m is matched to w.

A matching μ is *blocked by an individual* k if k prefers being single to being matched with $\mu(k)$, that is, $k > k\ \mu(k)$. A matching μ is *blocked by a pair of agents* (m,w) if they each prefer each other to the partner they receive at μ, that is, $w > m\ \mu(m)$ and $m > w\ \mu(w)$.

A matching μ is *stable* if it isn't blocked by any individual or pair of agents.

A stable matching is Pareto efficient, and in the core, and in this simple model the set of (pairwise) stable matchings equals the core.

Theorem 1 (Gale and Shapley 1962) A stable matching exists for every marriage market.

Gale and Shapley approached this problem from a purely theoretical perspective, but proved this theorem via a constructive algorithm of the kind that has subsequently turned up at the heart of a variety of clearing houses.

Deferred Acceptance Algorithm, with Men Proposing

(roughly the Gale and Shapley 1962 version).

Step 0. If some preferences are not strict, arbitrarily break ties (for example, if some m is indifferent between wi and wj, order them consecutively in alphabetical order. Different agents may break ties differently: that is, tie-breaking can be decentralized by having each agent fill out a strict preference list...).

Step 1(a). Each man m proposes to his 1st choice (if he has any acceptable choices).

Step1(b). Each woman rejects any unacceptable proposals and, if more than one acceptable proposal is received, 'holds' the most preferred and rejects all others. . . .

Step k(a). Any man who was rejected at step k-1 makes a new proposal to its most preferred acceptable mate who hasn't yet rejected him. (If no acceptable choices remain, he makes no proposal.)

Step k(b). Each woman holds her most preferred acceptable offer to date, and rejects the rest. STOP: when no further proposals are made, and match each woman to the man (if any) whose proposal she is holding.

Note that the proof of the theorem now follows from the observation that the matching produced in this way is itself stable. If some man would prefer to be matched to a woman other than his assigned mate, he must, according to the algorithm, have already proposed to her, and she has rejected him, meaning she has a man she strictly prefers, hence they cannot form a blocking pair.

Roth (1984) showed that the algorithm adopted by the medical clearing house in the 1950s was equivalent to the hospital proposing deferred acceptance algorithm. Gale and Shapley observed that which side of the market proposes in a deferred acceptance algorithm has consequences.

Theorem 2 (Gale and Shapley 1962) When all men and women have strict preferences, there always exists an M-optimal stable matching (that every man likes at least as well as any other stable matching), and a W-optimal stable matching. Furthermore, the matching μ M produced by the deferred acceptance algorithm with men proposing is the M-optimal stable matching. The W-optimal stable matching is the matching μ W produced by the algorithm when the women propose.

Note that the algorithm has been stated as if people take actions in the course of the algorithm, and we can ask whether those actions would best serve their interests. To put it another way, is it possible to design a clearing house in which a matching is produced from participants' stated rank order lists in such a way that it will never be in someone's interest to submit a rank order list different from their true preferences? The following theorem answers that question in the negative.

Theorem 3 Impossibility Theorem (Roth – see Roth and Sotomayor 1990) No stable matching mechanism exists for which stating the true preferences is a dominant strategy for every agent.

However it is possible to design the mechanism so that one side of the market can never do any better than to state their true preferences.

Theorem 4 (Dubins and Freedman, Roth – see Roth and Sotomayor 1990)

The mechanism that yields the M-optimal stable matching (in terms of the stated preferences) makes it a dominant strategy for each man to state his true preferences.

The conclusions of Theorems–3 also hold for a variety of related models (in which firms employ multiple workers, and wages are explicitly allowed to vary; see, for example, Shapley and Shubik 1971; Kelso and Crawford 1982, for notable early models of matching with money, and see Roth and Sotomayor 1990; Hatfield and Milgrom 2005). However, when we look at many-to-one matching models (in which firms employ multiple workers but workers seek just one job), we have to be careful. It turns out that no procedures exist that give firms a dominant strategy, but that a worker proposing deferred acceptance algorithm still

makes it a dominant strategy for workers to state their true preferences (see Roth and Sotomayor 1990 for more details and further references). (These results are closely connected to related results in auction theory; see in particular Hatfield and Milgrom 2005; Milgrom 2004.)

When the market for medical residents was redesigned Roth and Peranson (1999), a number of practical complications had to be dealt with, such as the fact that about 1000 graduates a year go through the match as couples who wish to be matched to nearby jobs, and hence have joint preferences over pairs of residency programmes. While this can cause the set of stable matchings to be empty, in practice this has not proved to be a significant problem (see also Roth 2002, on engineering aspects of economic design).

One-Sided Matching Models

Shapley and Scarf's 'House' Markets

Another basic model of matching markets was introduced by Shapley and Scarf (1974). They model a simple barter economy in which each one of n agents owns an indivisible good (which they call a house) and has preferences over all houses in the economy. Each agent has use for only one house and trade is only feasible in houses (that is, there is no money in their model). An allocation μ in this context is a matching of houses and agents so that each agent receives one and only one house. An exchange in this market does not need to be bilateral. An allocation μ is in the core if no coalition (including single agent coalitions) of agents can improve upon it (in the sense that all are weakly better off and at least one is strictly better off) by swapping their own houses. Shapley and Scarf attribute to David Gale the following *top trading cycles* algorithm (TTC) which can be used to find a core allocation for any housing market:

Step 1: Each agent points to the owner of her most preferred house (which could possibly be herself). Since there are finite number of agents there is at least one *cycle* (where a cycle is an ordered list (i_1, i_2, …, i_k) of agents with each agent pointing to the next agent in the list and agent *ik* pointing to agent i_1). In each cycle the implied exchange is carried out and the procedure is repeated with the remaining agents.

In general, at.

Step k: Each remaining agent points to the owner of her most preferred house among the remaining houses. There is at least one cycle. In each cycle the implied exchange is carried out and the procedure is repeated with the remaining agents.

The algorithm terminates when each agent receives a house.

Theorem 5 (Shapley and Scarf 1974) The TTC algorithm yields an allocation in the core for each housing market.

The core has some remarkable properties in the context of housing markets. The following propositions summarize the most notable of these results.

While exchange is feasible only in houses, a *competitive allocation* of a housing market can be defined via 'token money'. There is an important relation between the core and the competitive allocation for this very basic barter economy.

Theorem 6 (Roth and Postlewaite 1977) There is a unique allocation in the core (which can be obtained with the TTC algorithm) when agents have strict preferences over houses. Moreover the unique core allocation coincides with the unique competitive allocation.

Another remarkable feature of this model is that the top trading cycles mechanism makes it safe for agents to reveal their true preferences.

Theorem 7 (Roth 1982) The core as a mechanism is *strategy-proof* when agents have strict preferences over houses. That is, truth-telling is a dominant strategy for all agents in the preference revelation game in which TTC is applied to the stated preferences to produce an allocation.

Moreover, it is essentially the only mechanism that is strategy-proof among those that are Pareto efficient and *individually rational* (in the sense that an agent never receives a house inferior to her own).

Theorem 8 (Ma 1994) The core is the only mechanism that is Pareto efficient, individually rational and strategy-proof.

House Allocation Problems

Hylland and Zeckhauser (1979) introduced the *house allocation problem* which only differs from housing markets in *property rights*: There are n houses to be allocated for n agents where each agent has use for only one house and has strict preferences over all houses. Unlike in housing markets, no agent owns a specific house. The mechanism known as *random serial dictatorship* (RSD) is widely used in real-life allocation problems of this sort, such as assigning students to dormitory rooms.

Under RSD agents are randomly ordered (from a uniform distribution) in a list and the first agent in the list is assigned her top choice house, the next agent is assigned her top choice among the remaining houses, and so on. In addition to its popularity in practice, RSD has good incentive and efficiency properties.

Theorem 9 RSD is *ex post* Pareto efficient and strategy-proof.

Recall that the only difference between house allocation problems and housing markets is the initial property rights, and the core is very well-behaved in the context of the latter. This observation motivates the mechanism *core from random endowments* (CRE): randomly assign houses to agents with uniform distribution, interpret the resulting matching as the initial allocation of houses, and pick the core of the resulting housing market. It turns out, CRE is equivalent to RSD.

Theorem 10 (Abdulkadiroglu and Sönmez 1998) For any house allocation problem CRE and RSD yield the same lottery and hence they are equivalent mechanisms.

House Allocation with Existing Tenants

Housing markets and house allocation problems have very different property rights. The former is a pure private ownership economy where each house 'belongs' to a specific agent, whereas in the latter no strict subset of the grand coalition has claims on any house. Abdulkadiroglu and Sönmez (1999) introduced the following hybrid *house allocation with existing tenants* model. There are two kinds of agents: *existing tenants* each of whom owns a house, and *newcomers* none of whom has claims on a specific house. In addition to the *occupied houses* owned by existing tenants, there are also *vacant houses*. As in house allocation problems no specific person or group has claims on any vacant house. Suppose that the number of newcomers is equal to the number of vacant houses and hence the number of agents is equal to the number of houses. Agents have strict preferences over all houses and each existing tenant is allowed to keep her current house.

Abdulkadiroglu and Sönmez introduced the following *you request my house −1 get your turn* algorithm (YRMH–IGYT) which generalizes TTC as well as RSD. Under YRMH–IGYT, agents are randomly ordered in a line and initially only the vacant houses are *available.* The first agent in the line is assigned her top choice provided that it is either her own house or an available house (in which case her own house becomes available) and the process continues with the next agent in the line.

If, however, her top choice is an occupied house, the line is adjusted and the owner of the requested house is moved right in front of the requester. The process continues in a similar way with either the owner of the requested house getting assigned his own house or an available house (making his own house available), or otherwise his requesting an occupied house and upgrading its owner to the top of the line. When the process continues in a similar way there will either be a *cycle* of existing tenants (as in TTC) who can swap their own houses or a *chain* $(i_1, i_2, \ldots, i_k)$ of agents where agent i_1 is assigned an available house and each of the following agents is assigned the preceding agent's house.

The resulting mechanism inherits the attractive properties of its 'parents'.

Theorem 11 (Abdulkadiroglu and Sönmez 1999) The YRMH–IGYT mechanism is strategy-proof, *ex post* Pareto efficient, and individually rational (in the sense that no existing tenant receives a house inferior to her own).

Kidney Exchange

Living donors are an important source of kidneys for transplantation. But a patient with a willing living donor may not be able to receive a transplant because of a blood-type or immunologic incompatibility between her and her donor. Recently transplant centres around the world developed the possibility of *pairwise kidney exchange* in which two such pairs can exchange donors in case the donor in each pair is compatible with the patient in the other. Another interesting option is *indirect kidney exchange* in which the patient of an incompatible pair receives priority in the deceased donor waiting list if her incompatible donor donates a kidney to that waiting list. However, prior to 2004 only a very few exchanges had been accomplished, in large part because the market wasn't thick, and no databases were being maintained of incompatible patient–donor pairs. In an effort to organize kidney exchange on a larger scale, Roth et al. (2004) introduced the following *kidney exchange* model. There are a number of patients each with a (possibly) incompatible donor. For each patient a subset of donors can feasibly donate a kidney and the patient has strict preferences over these donors and his own donor (who may or may not be compatible with him). In addition to ranking all compatible donors, each patient also ranks a 'waiting list option' which represents trading his donor's kidney with a priority in the waitlist. An *allocation* in this context is a matching of patients and donors such that:

- each patient is matched with either a donor or the waiting list option, and
- each donor can be matched with at most one patient while the waiting list option may be matched with multiple patients.

(The donors who remain unmatched are offered to the waitlist in exchange for the equal number of priorities awarded by the allocation). We are only interested in *individually rational* allocations where patients receive neither a donor nor the waiting list option unless it is indicated to be at least as good as his donor's kidney. If the waiting list option is ranked inferior to his donor for a patient, that means the patient is not interested in such an exchange. As in the case of house allocation with existing tenants model, an allocation consists of cycles and chains where

- each patient in a cycle receives a kidney from the donor of the next patient in the cycle, and
- all but the last patient in a chain receive a kidney from the donor of the next patient in the chain whereas the last patient in the chain receives a priority in the waiting list.

If the waiting list option is infeasible, then the resulting problem is formally equivalent to a housing market and therefore has a unique allocation in the core which can be obtained via the TTC algorithm. In this simpler model an allocation (including the one in the core) consists of only cycles. When the waiting list option is feasible an allocation can also have chains (which are indirect exchanges and their more elaborate versions). In this more general model Roth et al. (2004) introduce a class of *top trading cycles and chains* (TTCC) algorithms each of which extend the TTC. Among these algorithms Roth et al. (2004) identify one that is Pareto efficient and strategy-proof:

Theorem 12 (Roth et al. 2004) There exists a TTCC mechanism that is Pareto efficient and strategy-proof.

In practice, as kidney exchanges have become organized on a larger scale in New England and elsewhere (see Roth et al. 2005a, b, 2007), there has been a focus, for logistical reasons, on cycles and chains that are relatively short, typically only involving exchanges among two or three patient–donor pairs.

The deferred acceptance algorithm (for two-sided markets) also has some uses in

one-sided allocation problems in which children are to be allocated to schools, if the schools, although not active strategic players, have priorities over students that need to be treated like preferences (Abdulkadiroglu and Sönmez 2003).

Design and Engineering

Introducing a Centralized Stable Match

Of the several dozen markets and submarkets we know of that established clearing houses in response to unravelling in a (two-sided) labour market, those that produce stable matchings have been most successful. Of particular note in this regard are the markets used in the various regions of the British National Health Service. In the 1960s, these markets suffered from the same kind of unravelling that had afflicted the American medical market in the 1940s. A Royal Commission recommended that each region organize a centralized clearing house (see Roth 1991), and the various regions each invented their own matching algorithms, some of which were stable and some of which were not (an example of such unstable algorithms will be given later). Those clearing houses that produced stable matches succeeded, while those that did not most often failed and were abandoned. But over a broad range of markets, the correlation between stability and success in halting unravelling isn't perfect; some unstable mechanisms remain in use, and some stable mechanism have occasionally failed, as we will discuss later. And there are other differences between markets than the way their clearing houses are designed. This is why, in order to establish that producing a stable outcome is an important feature for the success of a match, controlled experiments in the laboratory can be informative.

The laboratory experiments reported by Kagel and Roth (2002) help to verify the influence of a stable or unstable matching mechanism. After unravelling had begun in a small laboratory market, a clearing house was introduced using either the stable deferred acceptance algorithm or the unstable algorithms that failed in various regions of the British National Health service (Roth 1991). In the lab, as in the field, participants learned to wait for and use the stable algorithm, but learned to arrange their matches early and thus avoid using the unstable algorithm. Note that a laboratory market is quite different from a naturally occurring labour market, but it has the advantage that it allows the effect of the different algorithms to be observed in an environment in which everything else is the same.

Centralized clearing houses that yield stable outcomes have sometimes been introduced to organize markets suffering from failures other than unravelling (and the resulting lack of thickness), but related to congestion or the safety of revealing private information.

Examples of algorithms that produce unstable outcomes, but have been used in a number of market clearing houses, are so called priority algorithms, used for example by some British clearing houses, and also in several school choice problems in the United States. A priority algorithm classifies different matches in terms of priorities, based on the rank orders submitted, and then makes feasible matches in order of priority. In Boston, for example, the centralized system attempted to give as many students as possible their first choice school. The difficulty with the system was that students who did not get assigned to their first choice were much less likely to be assigned to the school they had listed as their second choice than they would have been if they had listed it as their first choice, since those schools often get filled by students who list them as their first choice. This means participants have strong incentives to not report their preferences truthfully, if there is a good chance that they would not be admitted to their true first choice school; it might be wiser to list their second-choice school as their first choice. The newly adopted Boston clearing house fixes this problem using a deferred acceptance algorithm (Abdulkadiroglu et al. 2005, 2006).

Some markets manage to halt unravelling, but still suffer from congestion. The market for clinical psychologists (before it reorganized through a modified deferred acceptance algorithm, see Roth and Xing 1997) and the match of students to New York City high schools before it was redesigned

(Abdulkadiroglu et al. 2005, 2006) are good examples. Clinical psychologists tried to run a deferred acceptance algorithm over the phone in the course of a day, 'match day,' from 9:00 a.m. to 4:00 p.m. All offers had to remain open until 4 p.m., and students were supposed to hold only one offer at a time. Even though turnaround time in this market was very fast (offers took about five minutes, rejections about one minute), simulating a deferred acceptance algorithm in real time, for a market with about 2000 positions in 500 programmes, takes much longer than the seven hours of match day. (And making the market longer may increase the effects of congestion, if it means that participants can no longer stay by the phone for the whole market, so that the time for an offer to be made and rejected becomes disproportionately longer.) Congestion is an issue whenever a large number of offers have to be made. The system used to assign students to New York City high schools used to be carried out through the mails, and over 30,000 students a year were 'stranded' on waiting lists and had to be assigned to a school for which they had expressed no preference. The new New York City clearing house is able to process preferences quickly, and in the four years following its adoption in 2003 fewer than 3000 students had to be assigned each year to a school for which they had expressed no preference.

What Are the Effects of a Centralized Match?

Centralized clearing houses can help make markets thick and uncongested, and avoid unravelling. Studying their effect on various markets can also help us understand how clearing houses and the timing of the market (for example, how far a labour market operates in advance of employment) influence the outcome of the market in other respects. For example, the market for gastroenterology fellows provides us with a natural case study of the effects of a clearing house not only on hiring practices (namely the timing of the market, and the kinds of offers that are made), but also employment opportunities, job placement and the potential impact on wages.

Gastroenterology fellows are doctors who have completed three years of residency in internal medicine, and are now employed in a fellowship that will result in their becoming board certified sub-specialists in gastroenterology. The market in which gastroenterology fellows are hired operated in a decentralized way for many years, and experienced the problems of congestion, unravelling and exploding offers, as described above in connection with the market for medical residents. In 1986, various internal medicine sub-specialties organized a clearing house called the Medical Specialties Matching Program (MSMP), sponsored by and organized along the same lines as the NRMP (which operates the resident match). But in the mid-1990s, gastroenterology fellowship programmes, and applicants, started to defect from the match, and the gastroenterology market again unravelled. A match was successfully re-established only in 2006 (Niederle et al. 2006). In those intervening years, as the market unravelled, the national market broke up into more local markets (Niederle and Roth 2003b). Fellowship programmes, particularly smaller ones, had a larger tendency to hire their own residents than under a centralized match.

A second aspect of the outcome that received prominence in 2002 is the question of whether a match affects wages. An antitrust lawsuit against the NRMP and numerous other defendants was brought in 2002 by 16 law firms on behalf of three former residents seeking to represent the class of all former residents (and naming as defendants a class including all hospitals that employ residents).

Niederle and Roth (2003a) showed empirically that in fact there is no difference in wages between medicine sub-specialties that use a match and those that don't. The suit was dismissed in 2004 following legislation intended to clarify that the medical match is a marketplace and does not violate antitrust laws.

The theory of the complaint was that a match holds down wages for residents and fellows. Bulow and Levin (2006) present a very stylized theoretical model providing some logical support for this possibility, by comparing a market with impersonal prices (to represent the NRMP) with perfectly competitive prices at which each worker is paid his or her marginal product. Subsequent

theoretical papers have shown that the conclusion about wage suppression doesn't necessarily follow if the model is expanded to include the possibility of firms hiring more than one worker (Kojima 2007), or when the model incorporates the actual procedures by which the medical match is conducted (Niederle 2007). Furthermore, decentralized markets may often fail to achieve stable outcomes (Niederle and Yariv 2007).

Beyond Centralized Matching. Why Do Some Markets Work Well, While Others Do Not?

We have seen that stability is an important feature for a centralized match to remain in use. However, the history of the gastroenterology market shows that producing a stable outcome is not sufficient to guarantee a successful clearing house. For a centralized match to work well, participants need to have incentives to participate in the match. McKinney et al. (2005) observed that the collapse of the gastroenterology fellowship match seems to have been caused by an unusual shock to the supply of highly qualified gastroenterology fellows, a kind of shock that was not observed in other internal medicine sub-specialties that continued to use a match. Furthermore, market conditions seemed to have stabilized, so that a centralized match would work well once again, if it could be successfully reinstated.

However, many gastroenterology fellowship programmes, when they considered reinstituting a match, were concerned that, while they were willing to refrain from making the early offers that had become customary, and wait for the match, their main competitors would continue to make early exploding offers to promising applicants. Such concerns could effectively prevent a successful restart of a centralized clearing house.

This raises the more general question as to why some markets unravel and experience congestion problems in the first place, while others do not. Empirically, most markets that experience congestion also experience that employers (hospitals, federal judges, colleges...) make short-term offers, with a binding deadline, and in which the acceptance of an offer is often effectively binding (Niederle and Roth 2007, for descriptions in the markets for law graduates, and for college admissions, see for example, Avery et al. 2001, 2003, 2007).

On the other hand there are markets that do not unravel, such as the market for graduate school admission. In this market, a policy (adopted by the large majority of universities) states that offers of admission and financial support to graduate students should remain open until 15 April. Furthermore, a student faced with an earlier deadline is explicitly encouraged to accept this offer, and, in case a better one is received before 15 April, to renege on that former acceptance. This of course makes early exploding offers much less attractive to make. Niederle and Roth (2007) explore environments in which either eliminating the possibility of making exploding offers or making early acceptances non-binding helps prevents markets from operating inefficiently early.

These insights were used to help reorganizing the gastroenterology fellowship match. To reduce the concerns of programmes that their competitors would start making exploding offers before the match, a resolution was adopted by the four main professional gastroenterology organizations that stated that acceptances made before the match were not to be considered binding, and such applicants could still change their minds and participate in the match. For an account of the effects of a centralized clearinghouse on the outcomes of a market, and the experience of the gastroenterology fellowship market, see Niederle and Roth (2008).

Directions for Future Research

As economists' understanding of the matching function of markets increases, and as economists are more often called upon to help design markets, one challenge will be to understand better how decentralized markets work well or badly, and not only in the final transactions.

For example, a common problem in many entry-level labour markets (and in dating and marriage markets) is that participants do not have well formed preferences over potential matching partners, and forming those preferences is often very costly. For example, in the American market for assistant professors, economics departments

receive hundreds of applications for any position, but in general interview only about 30 candidates at the annual winter meetings. From among those they interview, they must decide whom to fly out for extended campus visits and seminars, and it is from among this latter set of candidates that they eventually choose to whom to make an offer. Because this is a time-consuming and costly process many departments have to take care to interview applicants who not only have a good chance of being desirable colleagues, but who also have a good chance of accepting an offer if one is made. This often amounts to a coordination problem: not all departments should interview the same applicants. Allowing applicants to credibly submit information about their interest in particular schools can help alleviate this coordination problem, and in 2007 the American Economic Association implemented a signalling mechanism of this sort in the market for economists.

In general, the study of the matching function of markets has directed attention at the design of rules and procedures of both centralized and decentralized markets. The goal of the growing interest among economists in matching and market design is to understand the operation of markets, both centralized and decentralized, well enough so we can fix them when they're broken.

See Also

- ▶ Experimental Economics
- ▶ Experimental Labour Economics
- ▶ Game Theory
- ▶ Labour Market Institutions
- ▶ Matching
- ▶ Mechanism Design (New Developments)
- ▶ Mechanism Design Experiments

Bibliography

Abdulkadiroglu, A., and T. Sönmez. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66: 689–701.

Abdulkadiroglu, A., and T. Sönmez. 1999. House allocation with existing tenants. *Journal of Economic Theory* 88: 233–260.

Abdulkadiroglu, A., and T. Sönmez. 2003. School choice: A mechanism design approach. *American Economic Review* 93: 729–747.

Abdulkadiroglu, A., P.A. Pathak, and A.E. Roth. 2005a. The New York City high school match. *American Economic Review* 95: 364–367.

Abdulkadiroglu, A., P.A. Pathak, A.E. Roth, and T. Sönmez. 2005b. The Boston public school match. *American Economic Review* 95: 368–371.

Abdulkadiroglu, A., P.A. Pathak, and A.E. Roth. 2006a. *Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match* (Working paper). Harvard University.

Abdulkadiroglu, A., P.A. Pathak, A.E. Roth, and T. Sönmez. 2006b. *Changing the Boston school choice mechanism* (Working Paper No. 11965). Cambridge, MA: NBER.

Avery, C., C. Jolls, R.A. Posner, and A.E. Roth. 2001. The market for federal judicial law clerks. *University of Chicago Law Review* 68: 793–902.

Avery, C., A. Fairbanks, and R. Zeckhauser. 2003. *The early admissions game: Joining the elite*. Cambridge, MA: Harvard University Press.

Avery, C., C. Jolls, R.A. Posner, and A.E. Roth. 2007. The new market for federal judicial law clerks. *University of Chicago Law Review* 74: 447–486.

Bulow, J., and J. Levin. 2006. Matching and price competition. *American Economic Review* 96: 652–668.

Gale, D., and L.S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Hatfield, J., and P. Milgrom. 2005. Matching with contracts. *American Economic Review* 95: 913–935.

Hylland, A., and R. Zeckhauser. 1979. The efficient allocation of individuals to positions. *Journal of Political Economy* 87: 293–314.

Kagel, J.H., and A.E. Roth. 2000. The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment. *Quarterly Journal of Economics* 115: 201–235.

Kelso, A.S., and V.P. Crawford. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica* 50: 1483–1504.

Kojima, F. 2007. Matching and price competition. *American Economic Review* 97: 1027–1031.

Li, H., and S. Rosen. 1998. Unraveling in matching markets. *American Economic Review* 88: 371–387.

Ma, J. 1994. Strategy-proofness and the strict core in a market with indivisibilities. *International Journal of Game Theory* 23: 75–83.

McKinney, C.N., M. Niederle, and A.E. Roth. 2005. The collapse of a medical labor clearinghouse (and why such failures are rare). *American Economic Review* 95: 878–889.

Milgrom, P. 2004. *Putting auction theory to work*. Cambridge: Cambridge University Press.

Niederle, M. 2007. Competitive wages in a match with ordered contracts. *American Economic Review*.

Niederle, M., and A.E. Roth. 2003a. Relationship between wages and presence of a match in medical fellowships. *Journal of the American Medical Association* 290: 1153–1154.
Niederle, M., and A.E. Roth. 2003b. Unraveling reduces mobility in a labor market: Gastroenterology with and without a centralized match. *Journal of Political Economy* 111: 1342–1352.
Niederle, M., and A.E. Roth. 2007. *Making markets thick: How norms governing exploding offers affect market performance* (Working paper).
Niederle, M., and A.E. Roth. 2008. The effects of a central clearinghouse on job placement, wages, and hiring practices. In *Labor market intermediation*, ed. D. Autor. Chicago: University of Chicago Press.
Niederle, M., and L. Yariv. 2007. *Matching through decentralized markets* (Working paper).
Niederle, M., D.D. Proctor, and A.E. Roth. 2006. What will be needed for the new GI fellowship match to succeed? *Gastroenterology* 130: 218–224.
Roth, A.E. 1982. Incentive compatibility in a market with indivisible goods. *Economics Letters* 9: 127–132.
Roth, A.E. 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy* 92: 991–1016.
Roth, A.E. 1991. A natural experiment in the organization of entry level labor markets: Regional markets for new physicians and surgeons in the U.K. *American Economic Review* 81: 415–440.
Roth, A.E. 2002. The economist as engineer: Game theory, experimental economics and computation as tools of design economics. *Econometrica* 70: 1341–1378.
Roth, A.E., and E. Peranson. 1999. The redesign of the matching market for American physicians: Some engineering aspects of economic design. *American Economic Review* 89: 748–779.
Roth, A.E., and A. Postlewaite. 1977. Weak versus strong domination in a market with indivisible goods. *Journal of Mathematical Economics* 4: 131–137.
Roth, A.E., and M. Sotomayor. 1990. *Two-sided matching: A study in game – Theoretic modeling and analysis*. New York: Cambridge University Press.
Roth, A.E., and X. Xing. 1994. Jumping the gun: Imperfections and institutions related to the timing of market transactions. *American Economic Review* 84: 992–1044.
Roth, A.E., and X. Xing. 1997. Turnaround time and bottlenecks in market clearing: Decentralized matching in the market for clinical psychologists. *Journal of Political Economy* 105: 284–329.
Roth, A.E., T. Sönmez, and M.U. Ünver. 2004. Kidney exchange. *Quarterly Journal of Economics* 119: 457–488.
Roth, A.E., T. Sönmez, and M.U. Ünver. 2005a. A kidney exchange clearinghouse in New England. *American Economic Review* 95: 376–380.
Roth, A.E., T. Sönmez, and M.U. Ünver. 2005b. Pairwise kidney exchange. *Journal of Economic Theory* 125: 151–188.
Roth, A.E., T. Sönmez, M.U. Ünver, F.L. Delmonico, and S.L. Saidman. 2006. Utilizing list exchange and undirected Good Samaritan donation through 'chain' paired kidney donations. *American Journal of Transplantation* 6: 2694–2705.
Roth, A.E., T. Sönmez, and M.U. Ünver. 2007. Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences. *American Economic Review* 97: 828–851.
Shapley, L., and H. Scarf. 1974. On cores and indivisibility. *Journal of Mathematical Economics* 1: 23–37.
Shapley, L., and M. Shubik. 1971. The assignment game I: The core. *International Journal of Game Theory* 1: 111–130.

Matching Estimators

Petra E. Todd

Abstract

Matching methods are a popular method for evaluating the effects of programme or other treatment interventions. This article reviews recent developments in the econometric literature on matching estimators, including the assumptions required to justify their application, different ways of implementing the estimators and some recent empirical applications.

Keywords

Bootstrap; Curse of dimensionality; Kernel estimation; Local linear estimation; Matching; Matching estimators; Nearest-neighbour matching; Programme effect; Propensity score; Semiparametric estimation; Treatment effect

JEL Classifications

C13

Introduction

Matching is a widely used non-experimental method of evaluation that can be used to estimate

the average effect of a treatment or programme intervention. The method compares the outcomes of programme participants with those of matched non-participants, where matches are chosen on the basis of similarity in observed characteristics. One of the main advantages of matching estimators is that they typically do not require specifying the functional form of the outcome equation and are therefore not susceptible to misspecification bias along that dimension. Traditional matching estimators pair each programme participant with a single matched non-participant (see, for example, Rosenbaum and Rubin 1983), whereas more recently developed estimators pair programme participants with multiple non-participants and use weighted averaging to construct the matched outcomes.

We next define some notation and discuss how matching estimators solve the evaluation problem. Much of the treatment effect literature is built on the potential outcomes framework of Fisher (1935), exposited more recently in Rubin (1974) and Holland (1986). The framework assumes that there are two potential outcomes, denoted (Y_0, Y_1) that represent the states of being without and with treatment. An individual can be in only one state at a time, so only one of the outcomes is observed. The outcome that is not observed is termed a *counterfactual outcome*. The treatment impact for an individual is

$$\Delta = Y_1 - Y_0,$$

which is not directly observable. Assessing the impact of a programme intervention requires making an inference about what outcomes would have been observed in the no-programme state. Let $D = 1$ for persons who participate in the programme and $D = 0$ for persons who do not. The $D = 1$ sample often represents a select group of persons who were deemed eligible for a programme, applied to it, got accepted into it and decided to participate in it. The outcome that is observed is $Y = DY_1 + (1 - D)Y_0$.

Before considering different parameters of interest and their estimation, we first consider what is available directly from the data. The conditional distributions $F(Y_1|\, X, D = 1)$ and $F(Y_0|\, X, D = 0)$ can be recovered from the observations on Y_1 and Y_0, but not the joint distributions $F(Y_0, Y_1|\, X, D = 1)$, $F(Y_0, Y_1|\, X)$ or the impact distribution, $F(\Delta|\, X, D = 1)$. Because of this missing data problem, researchers often aim instead on recovering some features of the impact distribution, such as its mean. The parameter that is most commonly the focus of evaluation studies is the *mean impact of treatment on the treated*, $TT = E(Y_1 - Y_0|\, D = 1)$, which gives the benefit of the programme to programme participants. (If the outcome were earnings and the TT parameter exceeded the average cost of the programme, then the programme might be considered to at least cover its costs.)

Matching estimators typically assume that there exist a set of observed characteristics Z such that outcomes are independent of programme participation conditional on Z. That is, it is assumed that the outcomes (Y_0,Y_1) are independent of participation status D conditional on Z,

$$(Y_0, Y_1) \perp\!\!\!\perp D|Z. \tag{1}$$

The independence condition can be equivalently represented as $\Pr(D = 1|\, Y_0, Y_1, Z) = \Pr(D = 1|\, Z)$, or $E(D|\, Y_0, Y_1, Z) = E(D|\, Z)$. In the terminology of Rosenbaum and Rubin 1983, treatment assignment is 'strictly ignorable' given Z. It is also assumed that for all Z there is a positive probability of either participating ($D = 1$) or not participating ($D = 0$) in the programme: that is,

$$0 < \Pr(D = 1|\, Z) < 1. \tag{2}$$

This assumption is required so that matches for $D = 0$ and $D = 1$ observations can be found. If assumptions (1) and (2) are satisfied, then the problem of determining mean programme impacts can be solved by substituting the Y_0 distribution observed for matched on Z non-participants for the missing participant Y_0 distribution.

The above assumptions are overly strong if the parameter of interest is the mean impact of treatment on the treated (TT), in which case a weaker conditional mean independence assumption on Y_0 suffices (see Heckman et al. 1998a, b):

$$E(Y_0|Z,D=1) = E(Y_0|Z,D=0) = E(Y_0|Z). \quad (3)$$

Furthermore, when TT is the parameter of interest, the condition $0 < \Pr(D=1|Z)$ is also not required, because that condition is only needed to guarantee a participant analogue for each non-participant. The TT parameter requires only

$$\Pr(D=1|Z) < 1. \quad (4)$$

Under these assumptions, the mean impact of the programme on programme participants can be written as

$$\begin{aligned}\Delta_{TT} &= E(Y_1 - Y_0|D=1)\\ &= E(Y_1|D=1) - E_{Z|D=1}\{E_Y(Y|D=1,Z)\}\\ &= E(Y_1|D=1) - E_{Z|D=1}\{E_Y(Y|D=0,Z)\},\end{aligned}$$

where the second term can be estimated from the mean outcomes of the matched on Z comparison group. (The notation $E_{Z|D=1}$ denotes that the expectation is taken with respect to the $f(Z|D=1)$ density.)

Assumption (3) implies that D does not help predict values of Y_0 conditional on Z which rules out selection into the programme directly on values of Y_0. However, there is no similar restriction imposed on Y_1, so the method does allow individuals who expect to experience higher levels of Y_1 to select into the programme on the basis of that information. For estimating the *TT* parameter, matching methods allow selection into treatment to be based on possibly unobserved components of the anticipated programme impact, but only in so far as the programme participation decisions are based on the unobservable determinants of Y_1 and not those of Y_0.

Second, the matching method also requires that the distribution of the matching variables, Z, not be affected by whether the treatment is received. For example, age, gender, and race would generally be valid matching variables, but marital status may not be if it were potentially affected by receipt of the programme. To see why this assumption is necessary, consider the term

$$\begin{aligned}&E_{Z|D=1}\{E_Y(Y|D=0,Z)\}\\ &= \int_{z\in Z}\int_{y\in Y} y f(y|D=0,z) f(z|D=1)\mathrm{d}z.\end{aligned}$$

It uses the $f(z|D=1)$ conditional density to represent the density that would also have been observed in the no treatment ($D=0$) state, which rules out the possibility that receipt of treatment changes the density of Z. Variables that are likely to be affected by the treatment or programme intervention cannot be used in the set of matching variables.

With non-experimental data, there may or may not exist a set of observed conditioning variables for which (1) and (2), or (3) and (4), hold. A finding of Heckman et al. (1997) and Heckman et al. (1996, 1998a, b) in their application of matching methods to data from the Job Training and Partnership Act (JTPA) programme is that (2) and (4) were not satisfied, because no match could be found for a fraction of the participants. If there are regions where the support of Z does not overlap for the $D=1$ and $D=0$ groups, then matching is justified only when performed over the *region of common support*. The estimated treatment effect must then be defined conditionally on the region of overlap. Some methods for empirically determining the overlap region are described below.

Matching estimators can be difficult to implement when the set of conditioning variables Z is large. If Z are discrete, small-cell problems may arise. If Z are continuous and the conditional mean $E(Y_0|D=0,Z)$ is estimated nonparametrically, then convergence rates will be slow due to the so-called *curse of dimensionality* problem. Rosenbaum and Rubin (1983) provide a theorem that can be used to address this dimensionality problem. They show that for random variables Y and Z and a discrete random variable D

$$\begin{aligned}&E(D|Y,P(D=1|Z))\\ &\quad = E(E(D|Y,Z)|Y,\Pr(D=1|Z)),\end{aligned}$$

so that

$$\begin{aligned}&E(D|Y,Z) = E(D|Z)\\ &\Rightarrow E(D|Y,\Pr(D=1|Z) = E(D|\Pr(D=1|Z)).\end{aligned}$$

This result implies that, when Y_0 outcomes are independent of programme participation conditional on Z, they are also independent of participation conditional on the probability of participation, $P(Z) = \Pr(D = 1|\, Z)$. That is, when matching on Z is valid, matching on the summary statistic $\Pr(D = 1|\, Z)$ (the *propensity score*) is also valid. Provided that $P(Z)$ can be estimated parametrically (or semiparametrically at a rate faster than the nonparametric rate), matching on the propensity score reduces the dimensionality of the matching problem to that of a univariate problem. For this reason, much of the literature on matching focuses on propensity score matching methods. (Heckman et al. 1998a, b, and Hahn 1998, consider whether it is better in terms of efficiency to match on $P(X)$ or on X directly.) With the use of the Rosenbaum and Rubin (1983) theorem, the matching procedure can be broken down into two stages. In the first stage, the propensity score $\Pr(D = 1|\, Z)$ is estimated, using a binary discrete choice model. (Options for first the stage estimation include, for example, a parametric logit or probit model or a semiparametric estimator, such as semiparametric least squares – Ichimura 1993 – maximum score – Manski 1973 – smoothed maximum score – Horowitz 1992 – or semiparametric maximum likelihood – Klein and Spady 1993. If $P(Z)$ were estimated using a fully nonparametric method, then the curse of dimensionality problem would reappear.) In the second stage, individuals are matched on the basis of their predicted probabilities of participation.

We next describe a simple model of the programme participation decision to illustrate the kinds of assumptions needed to justify matching. (This model is similar to an example given in Heckman et al. 1999.) Assume that an individual chooses whether to apply to a training programme on the basis of the expected benefits. He or she compares the expected earnings streams with and without participating, taking into account opportunity costs and net of some random training cost ε, which may include a psychic component expressed in monetary terms. The participation decision is made at time $t = 0$ and the training programme lasts for periods 1 through τ, during which time earnings are zero. The information set used to determine expected earnings is denoted by W, which might include, for example, earnings and employment history. The participation model is

$$D = 1 \text{ if } E\left(\sum_{j=\tau}^{T} \frac{Y_{1j}}{(1+r)^j} - \sum_{k=1}^{T} \frac{Y_{0k}}{(1+r)^k} \,|\, W\right) > \varepsilon + Y_{00}, \text{ else } D = 0.$$

The terms in the right-hand side of the inequality are assumed to be known to the individual but not to the econometrician.

$$\begin{aligned} &\text{If} f(Y_{0k}|\, \varepsilon + Y_{00}, X) = f(Y_{0k}|\, X), \\ &\quad \text{then } E(Y_{0k}|\, X, D = 1) \\ &\quad = E(Y_{0k}|\, X, \varepsilon + Y_{00} < \eta(W)) \\ &\quad = E(Y_{0k}|\, X), \end{aligned}$$

which would justify application of a matching estimator. This assumption places restrictions on the correlation structure of the earnings residuals. For example, the assumption would not be plausible if $X = W$ and $Y_{00} = Y_{0k}$, because knowing that a person selected into the programme ($D = 1$) would likely be informative about subsequent earnings. We could assume, however, a model for earnings

$$Y_{0k} = \varphi(X) + v_{0k},$$

such as where v_{0k} follows an MA(q) process with $q < k$, which would imply that Y_{0k} and Y_{00} are uncorrelated conditional on X. The matching method does not require that everything in the information set be known, but it does assume sufficient information to make the selection on observables assumption plausible.

Cross-Sectional Matching Methods

For notational simplicity, let $P = P(Z)$. A prototypical propensity score matching estimator takes the form

$$\begin{aligned} \hat{\alpha}_M &= \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left[Y_{1i} - \hat{E}(Y_{0i}|\, D = 1, P_i)\right] \\ \hat{E}(Y_{0i}|\, D = 1, P_i) &= \sum_{j \in I_0} W(i,j) Y_{0j}, \end{aligned} \tag{5}$$

where I_1 denotes the set of programme participants, I_0 the set of non-participants, S_P the region of common support (see below for ways of constructing this set). n_1 is the number of persons in the set $I_1 \cap S_P$. The match for each participant $I \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of non-participants, where the weights $W(i, j)$ depend on the distance between P_i and P_j. Define a neighbourhood $C(P_i)$ for each i in the participant sample. Neighbours for i are non-participants $j \in I_0$ for whom $P_j \in C(P_i)$. The persons matched to i are those people in set A_i where $A_i = \{j \in I_0 | P_j \in C(P_i)\}$. We describe a number of alternative matching estimators below, that differ in how the neighbourhood is defined and in how the weights $W(i, j)$ are constructed.

Alternative Ways of Constructing Matched Outcomes

Nearest-Neighbour Matching

Traditional, pairwise matching, also called *nearest-neighbour matching*, sets:

$$C(P_i) = \min_j \left\| P_i - P_j \right\|, j \in I_0.$$

That is, the non-participant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set. The estimator can be implemented either matching with or without replacement. When matching is performed with replacement, the same comparison group observation can be used repeatedly as a match. A drawback of matching without replacement is that the final estimate will usually depend on the initial ordering of the treated observations for which the matches were selected.

Caliper matching (Cochran and Rubin 1973) is a variation of nearest neighbour matching that attempts to avoid 'bad' matches (those for which P_j is far from P_i) by imposing a tolerance on the maximum distance $\|P_i - P_j\|$allowed. That is, a match for person i is selected only if $\|P_i - P_j\| < \varepsilon$, $j \in I_0$, where ε is a pre-specified tolerance. Treated persons for whom no matches can be found within the caliper are excluded from the analysis, which is one way of imposing a common support condition. A drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

Stratification or Interval Matching

In this variant of matching, the common support of P is partitioned into a set of intervals, and average treatment impacts are calculating through simple averaging within each interval. A weighted average of the interval impact estimates, using the fraction of the $D = 1$ population in each interval for the weights, provides an overall average impact estimate. Implementing this method requires a decision on how wide the intervals should be. Dehejia and Wahba (1999) implement interval matching using intervals that are selected such that the mean values of the estimated P_i and P_j are not statistically different from each other within intervals.

Kernel and Local Linear Matching

More recently developed matching estimators construct a match for each programme participant using a weighted average over multiple persons in the comparison group. Consider, for example, the nonparametric *kernel matching estimator*, given by

$$\hat{\alpha}_{KM} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)} \right\}$$

where $G(\ )$ is a kernel function and an is a bandwidth parameter. (See Heckman et al. 1997a, b, 1998a, b and Heckman et al., 1998a, b.) In terms of Eq. (5), the weighting function, $W(i,j)$, is equal to

$$\frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}.$$

For a kernel function bounded between -1 and 1, the neighbourhood is

$$C(P_i) = \left\{ \left| \frac{P_i - P_j}{a_n} \right| \leq 1 \right\}, j \in I_0.$$

M

Under standard conditions on the bandwidth and kernel,

$$\frac{\sum_{j\in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k\in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$$

is a consistent estimator of $E(Y_0|\ D = 1, P_i)$. (Specifically, we require that $G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$: One example of a kernel function is the quartic kernel, given by $G(s) = \frac{15}{16}(s^2 - 1)^2$ if $|s| < 1$, $G(s) = 0$ otherwise.)

Heckman et al. (1997) also propose a generalized version of kernel matching, called local linear matching. Recent research by Fan 1992a, b, demonstrated advantages of local linear estimation over more standard kernel estimation methods. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities; see Fan 1992a, b.) The local linear weighting function is given by

$$W(i,j) = \frac{G_{ij}\sum_{k\in I_0} G_{ik}(P_k - P_i)^2 - \left[G_{ij}(P_j - P_i)\right]\left[\sum_{k\in I_0} G_{ik}(P_k - P_i)\right]}{\sum_{j\in I_0} G_{ij}\sum_{k\in I_0} G_{ij}(P_k - P_i)^2 - \left(\sum_{k\in I_0} G_{ik}(P_k - P_i)\right)^2}. \tag{6}$$

As demonstrated in research by Fan (1992a, b), local linear estimation has some advantages over standard kernel estimation. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities (see Fan 1992a, b). Thus, local linear regression would be expected to perform better than kernel estimation in cases where the non-participant observations on P fall on one side of the participant observations.

To implement the matching estimator given by Eq. (5), the region of common support S_P needs to be determined. The common support region can be estimated by

$$\hat{S}_P = \left\{P : \hat{f}(P|D=1) > 0 \text{ and} \hat{f}(P|D=0) > c_q\right\},$$

where $\hat{f}(P|D=d), d \in \{0,1\}$ are standard non-parametric density estimators. To ensure that the densities are strictly greater than zero, it is required that the densities be strictly positive (that is, exceed zero by a certain amount), determined using a 'trimming level' q. That is, after excluding any P points for which the estimated density is zero, an additional small percentage of the remaining P points is excluded for which the estimated density is positive but very low. The set of eligible matches is thus given by

$$\hat{S}_q = \left\{P \in \hat{S}_P : \hat{f}(P|D=1) > c_q \text{ and} \hat{f}(P|D=0) > c_q\right\},$$

where c_q is the density cut-off level that satisfies

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i\in I_1 \cap \hat{S}\}} \left\{1\left(\hat{f}(P|D=1)\right) < c_q + 1\left(1\left(\hat{f}(P|D=0)\right) < c_q\right)\right\} \le q.$$

Here, J is the cardinality of the set of observed values of P that lie in $I_1 \cap \hat{S}_P$. That is, matches are constructed only for the programme participants for which the propensity scores lie in $\hat{S}_q$.

The above estimators are representations of matching estimators and are commonly used. They can be easily adapted to estimate other parameters of interest, such as the average effect of treatment on the untreated (UT = $E(Y_1 - Y_0|\ D = 0, X)$), or the average treatment effect (ATE = $E(Y_1 - Y_0|\ X)$), which is just a weighted average of treatment on the treated (TT) and treatment on the untreated (UT).

The recent literature has also developed alternative matching estimators that employ different weighting schemes to increase efficiency. See, for example, Hahn (1998) and Hirano et al. (2003) for

estimators that attain the semiparametric efficiency bound. The methods are not described in detail here, because those studies focus on the ATE and not on the average effect of treatment on the treated (TT) parameter. Heckman, Ichimura and Todd (1998) develop a regression-adjusted version of the matching estimator, which replaces Y_{0j} as the dependent variable with the residual from a regression of Y_{0j} on a vector of exogenous covariates. The estimator uses a Robinson (1988) type estimation approach to incorporate exclusion restrictions: that is, that some of the conditioning variables in an equation for the outcomes do not enter into the participation equation or vice versa. In principle, imposing exclusion restrictions can increase efficiency. In practice, though, researchers have not observed much gain from using the regression-adjusted matching estimator. Some alternatives to propensity score matching are discussed in Diamond and Sekhon (2005).

When Does Bias Arise in Matching?

The success of a matching estimator depends on the availability of observable data to construct the conditioning set Z, such that (1) and (2) are satisfied. Suppose only a subset $Z_0 \subset Z$ of the required variables is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha'_M = E_{P(Z_0)|D=1}(E(Y_1|P(Z_0), D=1) - E(Y_0|P(Z_0), D=0)). \quad (7)$$

The bias for the parameter of interest, $E(Y_1 - Y_0|\ D=1)$, is

$$\text{bias}_M = E(Y_0|D=1) - E_{P(Z_0)|D=1}\{E(Y_0|P(Z_0), D=0)\}.$$

There is no way of a priori choosing the set of Z variables to satisfy the matching condition or of testing whether a particular set meets the requirements. In rare cases, where data are available on a randomized social experiment, it is sometimes possible to ascertain the bias (see, for example, Heckman et.al 1997a, b; Dehejia and Wahba 1999, 2002; Smith and Todd 2005).

Difference-in-Difference Matching Estimators

The estimators described above assume that, after conditioning on a set of observable characteristics, outcomes are conditionally mean independent of programme participation. However, for a variety of reasons there may be systematic differences between participant and non-participant outcomes, even after conditioning on observables, which could lead to a violation of the identification conditions required for matching. Such differences may arise, for example, because of programme selectivity on unmeasured characteristics or because of levels differences in outcomes that might arise when participants and non-participants reside in different local labour markets or if the survey questionnaires used to gather the data differ in some ways across groups.

A difference-in-differences (DID) matching strategy, as defined in Heckman et al. (1997) and Heckman et al. (1998a, b), allows for temporally invariant differences in outcomes between participants and non-participants. This type of estimator matches on the basis of differences in outcomes using the same weighting functions described above. The propensity score DID matching estimator requires that

$$E(Y_{0t} - Y_{0t'}|P, D=1) = E(Y_{0t} - Y_{0t'}|P, D=0),$$

where t and t' are time periods after and before the programme enrolment date. This estimator also requires the support condition given above, which must now hold in both periods t and t'. The local linear difference-in-difference estimator is given by

$$\hat{\alpha}_{DM} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_P} W(i,j)(Y_{0tj} - Y_{0t'j}) \right\},$$

where the weights correspond to the local linear weights defined above. If repeated cross-section data are available, instead of longitudinal data, the estimator can be implemented as

M

$$\hat{\alpha}_{DM} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap S_P} \left\{ \left(Y_{1ti} - \sum_{j \in I_{0t} \cap S_P} W(i,j) Y_{0tj} \right\} - \frac{1}{n_{1t'}} \sum_{i \in I_{1t'} \cap S_P} \left\{ \left(Y_{1t'i} - \sum_{j \in I_{0t'}} W(i,j) Y_{0t'j} \right\},$$

where $I_{1t}, I_{1t'}, I_{0t}, I_{0t'}$ denote the treatment and comparison group data-sets in each time period.

Finally, the DID matching estimator allows selection into the programme to be based on anticipated gains from the programme in the sense that D can help predict the value of Y_1 given P. However, the method assumes that D does not help predict changes $Y_{0t} - Y_{0t'}$ conditional on a set of observables (Z) used in estimating the propensity score. In their analysis of the effectiveness of matching estimators, Smith and Todd (2005) found difference-in-difference matching estimators to perform much better than cross-sectional methods in cases where participants and non-participants were drawn from different regional labour markets and/or were given different survey questionnaires.

Matching When the Data are Choice-Based Sampled

The samples used in evaluating the impacts of programmes are often choice-based, with programme participants oversampled relative to their frequency in the population of persons eligible for the programme. Under choice-based sampling, weights are generally required to consistently estimate the probabilities of programme participation. (See, for example, Manski and Lerman 1977, for discussion of weighting for logistic regressions.) When the weights are unknown, Heckman and Todd (1995) show that with a slight modification matching methods can still be applied, because the odds ratio $(P/(1-P))$ estimated using a logistic model with incorrect weights (that is, ignoring the fact that samples are choice-based) is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores. Therefore, matching can proceed on the (misweighted) estimate of the odds ratio (or on the log odds ratio).

Using Balancing Tests to Check the Specification of the Propensity Score Model

As described earlier, the propensity score matching estimator requires the outcome variable to be mean independent of the treatment indicator conditional on the propensity score, $P(Z)$. An important consideration in implementation is how to choose Z. Unfortunately, there is no theoretical basis for choosing a particular set Z to satisfy the identifying assumptions, and the set is not necessarily the most inclusive one.

To guide in the selection of Z, there is some accumulated empirical evidence on how bias estimates depended on the choice of Z in particular applications. For example, Heckman et al. (1998a, b), Heckman et al. (1997) and Lechner (2001) show that the choice of variables included in Z can make a substantial difference to the estimator's performance. These papers found that biases tended to be higher when the participation equation was estimated using a cruder set of conditioning variables. One approach adopted is to select the set Z to maximize the percentage of people correctly classified under the model. Another finding in these papers is that the matching estimators performed best when the treatment and control groups were located in the same geographic area and when the same survey instrument was administered to both treatments and controls to ensure comparable measurement of outcomes.

Rosenbaum and Rubin (1983) suggest a method to aid in the specification of the propensity score model. The method does not provide guidance in choosing which variables to include in Z, but can help to determine which interactions and higher-order terms to include in the model for a given Z set. They note that for the true propensity score, the following holds:

$$Z \perp\!\!\!\perp D | \Pr(D = 1 | Z),$$

or equivalently $E(D|\ Z; \Pr(D = 1|\ Z)) = E(D|\ \Pr(D = 1|\ Z))$. The basic intuition is that, after conditioning on $\Pr(D = 1|Z)$, additional conditioning on Z should not provide new information about D. If after conditioning on the estimated

values of $P(D = 1|Z)$ there is still dependence on Z, this suggests misspecification in the model used to estimate $\Pr(D = 1|Z)$. The theorem holds for any Z, including sets Z that do not satisfy the conditional independence condition required to justify matching. As such, the theorem is not informative about what set of variables to include in Z.

This result motivates a specification test for $\Pr(D = 1|Z)$, that is a test whether or not there are differences in Z between the $D = 1$ and $D = 0$ groups after conditioning on $P(Z)$. The test has been implemented in the literature a number of ways (see, for example Eichler and Lechner 2002; Dehijia and Wahba 1999, 2002; Smith and Todd 2005; Diamond and Sekohn 2005).

Assessing the Variability of Matching Estimators

The distribution theory for the cross-sectional and difference-in-difference kernel and local linear matching estimators described above is derived in Heckman et al. (1998). However, implementing the asymptotic standard error formulae can be cumbersome, so standard errors for matching estimators are often instead generating using bootstrap resampling methods. (See Efron and Tibshirani 1993, for an introduction to bootstrap methods, and Horowitz 2003, for a recent survey of bootstrapping in econometrics.) A recent paper by Abadie and Imbens (2006a) shows that standard bootstrap resampling methods are not valid for assessing the variability of nearest neighbour estimators, but can be applied to assess the variability of kernel or local linear matching estimators for a suitably chosen bandwidth. Abadie and Imbens (2006b) present alternative standard error formulae for assessing the variability of nearest neighbour matching estimators.

Applications

There have been numerous evaluations of matching estimators in recent decades. For a survey of many applications in the context of evaluating the effects of labour market programmes (see Heckman et al. 1999). More recently, propensity score matching estimators have been used in evaluating the impacts of a variety of programme interventions in developing countries. Jalan and Ravallion (1999) assess the impact of a workfare programme in Argentina (the *Trabajar* programme), and Jalan and Ravallion (2003) study the effects of public investments in piped water on child health outcomes in rural India. Galiani et al. (2005) use difference-in-difference matching methods to analyse the effects of privatization of water services on child mortality in Argentina. Other applications include Gertler et al. (2004) in a study of the effects of parental death on child outcomes, Lavy (2004) in a study of the effects of a teacher incentive programme in Israel on student performance, Angrist and Lavy (2001) in a study of the effects of teacher training on children's test scores in Israel, and Chen and Ravallion (2003) in a study of a poverty reduction project in China.

Behrman et al. (2004) use a modified version of a propensity score matching estimator to evaluate the effects of a preschool programme in Bolivia on child health and cognitive outcomes. They identify programme effects by comparing children with different lengths of duration in the programme, using matching to control for selectivity into alternative durations. Also, see Imbens (2000) and Hirano and Imbens (2004) for an analysis of the role of the propensity score with continuous treatments. Lechner (2001) extends propensity score analysis for the case of multiple treatments.

See Also

- ▶ Propensity Score
- ▶ Selection Bias and Self-Selection
- ▶ Semiparametric Estimation
- ▶ Treatment Effect

Bibliography

Abadie, A. and G. Imbens 2006a. On the failure of the bootstrap for matching estimators. Technical working paper no. 325. Cambridge, MA: NBER.

M

Abadie, A., and G. Imbens. 2006b. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74: 235–267.

Angrist, J., and V. Lavy. 2001. Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools. *Journal of Labor Economics* 19: 343–369.

Behrman, J., Y. Cheng, and P. Todd. 2004. Evaluating preschool programs when length of exposure to the program varies: A nonparametric approach. *The Review of Economics and Statistics* 86: 108–132.

Chen, S. and M. Ravallion 2003. Hidden impact? ex-post evaluation of an antipoverty program. Policy Research Working paper no. 3049. Washington, DC: World Bank.

Cochran, W., and D. Rubin. 1973. Controlling bias in observational studies. *Sankyha* 35: 417–446.

Dehejia, R., and S. Wahba. 1999. Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94: 1053–1062.

Dehejia, R., and S. Wahba. 2002. Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics* 84: 151–161.

Diamond, A. and J.S. Sekhon 2005. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Working paper, Department of Political Science, Berkeley.

Efron, B., and R. Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman and Hall.

Eichler, M., and M. Lechner. 2002. An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics* 9: 143–186.

Fan, J. 1992a. Design adaptive nonparametric regression. *Journal of the American Statistical Association* 87: 998–1004.

Fan, J. 1992b. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics* 21: 196–216.

Fisher, R.A. 1935. *Design of experiments*. New York: Hafner.

Friedlander, D., and P. Robins. 1995. Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review* 85: 923–937.

Galiani, S., P. Gertler, and E. Schargrodsky. 2005. Water for life: The impact of the privatization of water services on child mortality in argentina. *Journal of Political Economy* 113: 83–120.

Gertler, P., D. Levine, and M. Ames. 2004. Schooling and parental death. *The Review of Economics and Statistics* 86: 211–225.

Hahn, J. 1998. On the role of the propensity score in efficient estimation of average treatment effects. *Econometrica* 66: 315–331.

Heckman, J. and P. Todd 1995. Adapting propensity score matching and selection models to choice-based samples. Manuscript, Department of Economics, University of Chicago.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1996. Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proceedings of the National Academy of Sciences* 93: 13416–13420.

Heckman, J., J. Smith, and N. Clements. 1997a. Making the most out of social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies* 64: 487–536.

Heckman, J., H. Ichimura, and P. Todd. 1997b. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies* 64: 605–654.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998a. Characterizing selection bias using experimental data. *Econometrica* 66: 1017–1098.

Heckman, J., H. Ichimura, and P. Todd. 1998b. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65: 261–294.

Heckman, J., R. Lalonde, and J. Smith. 1999. The economics and econometrics of active labor market programs. In *Handbook of labor economics*, ed. O. Ashenfelter and D. Card, Vol. 3A. Amsterdam: North-Holland.

Hirano, K., and G. Imbens. 2004. The propensity score with continuous treatments. In *Applied bayesian modeling and causal inference from incomplete data perspectives*, ed. A. Gelman and X.L. Meng. New York: Wiley.

Hirano, K., G. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71: 1161–1189.

Holland, P.W. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81: 945–970.

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60: 505–532.

Horowitz, J.L. 2003. The bootstrap. In *Handbook of econometrics*, ed. J.J. Heckman and E.E. Leamer, Vol. 5. Amsterdam: North-Holland.

Ichimura, H. 1993. Semiparametric least squares and weighted SLS estimation of single index models. *Journal of Econometrics* 58: 71–120.

Imbens, G. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87: 706–710.

Jalan, J. and M. Ravallion 1999. Efficient estimation of average treatment effects: Evidence for argentina's trabajar program. Policy research working paper. Washington, DC: World Bank.

Jalan, J., and M. Ravallion. 2003. Does piped water reduce diarrhea for children in rural India. *Journal of Econometrics* 112: 153–173.

Klein, R.W., and R.H. Spady. 1993. An efficient semiparametric estimator for binary response models. *Econometrica* 61: 387–422.

LaLonde, R. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76: 604–620.

Lavy, V. 2002. Evaluating the effects of teachers' group performance incentives on pupil achievement. *Journal of Political Economy* 110: 1286–1387.
Lavy, V. 2004. Performance pay and teachers' effort, productivity and grading ethics. Working paper no. 10622. Cambridge, MA: NBER.
Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluations of active labor market policies in Europe*, ed. M. Lechner and F. Pfeiffer. Heidelberg: Physica.
Manski, C. 1973. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3: 205–228.
Manski, C., and S. Lerman. 1977. The estimation of choice probabilities from choice-based samples. *Econometrica* 45: 1977–1988.
Robinson, P. 1988. Root-*N* consistent nonparametric regression. *Econometrica* 56: 931–954.
Rosenbaum, P., and D. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.
Rosenbaum, P., and D. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39: 33–38.
Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.
Silverman, B.W. 1986. *Density estimation for statistics and data analysis*. London: Chapman and Hall.
Smith, J., and P. Todd. 2005. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics* 125: 305–353.

Matching Models: Empirics

Jeremy T. Fox

Abstract

A matching model takes a set of payoffs or outputs for all possible matches and produces a set of matches where no couple would prefer to deviate and become matched, instead of their assigned matches. Matching models are increasingly being estimated in empirical work in industrial organization, labour economics, public economics, and other fields. This article surveys methods for and applications of structural estimation for two-sided matching games.

Keywords

Cooperative games; Econometrics; Family economics; Industrial organization; Labour market; Marriage market; Matching models; Mergers; Pairwise stability; Structural estimation; Two-sided matching games

JEL Classifications

C78; C25; C35; G; H; J; L; O; R

Economists often observe data on relationships. We see who is in a relationship with whom: which firms merged with each other, which men are married to which women, and which bidders won which items in an auction. A matching model or matching game is one theoretical framework for modelling the equilibrium formation of these relationships. A relationship is termed a match. A matching model takes a set of payoffs or outputs for all possible matches and produces a set of matches where no couple would prefer to deviate and become matched, instead of their assigned matches. The robustness of the equilibrium to deviation by any potential couple suggests 'pairwise stability' as the term for an equilibrium.

The key economic idea in matching models is the rivalry to match with the most attractive partners. In marriage, men compete with each other to marry the most attractive women while women compete with each other to marry the most attractive men. There is scarcity on both sides of the market.

This entry focuses on research that formally estimates the parameters of a matching model using data on who matches with whom. Economists adopting this structural approach typically observe data on who matches with whom as well as exogenous characteristics of each agent. For example, economists studying marriage will observe the race, schooling level, religion, physical attractiveness and wage of each man and each women. The data also record which men married which women. Economists are willing to assume that the data represent a pairwise stable outcome to a particular matching game.

The structural approach means that researchers impose the structural model and estimate unknown parameters in the model. The advantages of the structural approach apply to more economic situations than just matching models, and have been explored elsewhere (Reiss and Wolak 2007). A quick summary is that the structural approach allows the computation of counterfactuals and the measurement of economic parameters that cannot be directly observed. Using the example of marriage, one type of counterfactual would be to explore how the equilibrium set of matches is altered as demographics change. Measurement is also important: if men and women each have several characteristics, how important are each of the characteristics in the payoffs to a match?

This article focuses on the use of matching games in structural empirical work. Other literatures have focused on centralized market design (for example the medical resident matching programme in the United States) and the descriptive interpretation of matching patterns.

Two-sided Matching Games

Most but not all empirical work has focused on two-sided matching games: agents can be divided into two sides, say men and women. Roth and Sotomayor (1990) is a useful text that explains simple matching games.

For the purposes of this article, I divide two-sided matching games into models with and without transfers. Gale and Shapley (1962) introduced the model where agents do not exchange money: men have preferences over women and women have preferences over men. Generically there will be a lattice of multiple pairwise stable outcomes in this model. Koopmans and Beckmann (1957), Shapley and Shubik (1972), and Becker (1973) study models where matched agents can exchange money and where agents have transferable utility. For one-to-one matching games such as marriage, generically there will be one set of pairwise stable physical matches in these models; there may be a continuum of transfers that support these physical matches. The choice of modelling framework depends on the researcher's understanding of the market in question.

The above papers allow each man, say, to marry at most one woman. There are extensions to many-to-one and many-to-many two-sided matching games. Complementarities between multiple matches involving the same agent are key to some of the empirical applications below (Fox 2009a; Fox and Bajari 2009). There are also one-sided and many-sided matching games.

There are more general matching games where other contract elements, such as the hours of work in a labour market, are determined as part of the pairwise stable outcome (Crawford and Knoer 1981; Kelso and Crawford 1982; Hatfield and Milgrom 2005). Matching games are mathematically linked to hedonic equilibrium models, although I will not explore the link here (Rosen 1974). There is also a clear link to models of frictions, such as search models, that also have observed agent heterogeneity (Shimer and Smith 2000; Atakan 2006).

Estimation Methods

Matching games share many similarities with the literature on estimating static, discrete Nash games, such as the well-known entry models of Berry (1992) and Bresnahan and Reiss (1991).

Matching games use pairwise stability and not Nash equilibrium, but many estimation challenges are similar. A key difficulty in matching games is that the number of agents in a market can be in the hundreds or thousands, compared to the three or four firms deciding to enter a market in some entry applications. The number of agents in matching empirical applications can make some estimators computationally infeasible.

Nested Solution Methods

The most straightforward way to estimate a matching game is to use simulated maximum likelihood or the simulated method of moments. These estimators require solving the model a fixed number of times for each iteration of an outer optimization routine. Simulation estimators are

conceptually straightforward but computationally burdensome.

Boyd and colleagues (2003) use the simulated method of moments to study the matching of public school teachers to schools in New York state. They use data on wages and assume the wages are exogenously determined. Their model without endogenous transfers has multiple stable matches, and they use the lattice structure of the equilibria to impose an equilibrium selection rule in estimation.

Full Likelihood Methods

In many cases, the full likelihood can be written down. In a study of the matching of venture capitalists to entrepreneurs, Sørensen (2007) uses an augmented likelihood approach where the unobserved payoffs of each match are treated as nuisance parameters and integrated out using a blocking structure in a Markov Chain Monte Carlo (MCMC) procedure. Sørensen does not use endogenous transfers and hence imposes an aligned preferences assumption that he proves generates a unique pairwise stable outcome.

The full likelihood approach can be computationally intractable in large matching markets. In a study of the matching between investment banks and firms undertaking an initial public offering, Akkus (2008) shows that the likelihood simplifies if the values of realized matches are recorded in the data. By using data on the payoffs of matches, estimation becomes easier.

Inequality Methods

With an application to automotive suppliers and automotive assemblers, Fox (2009a) introduces a maximum score estimator to estimate a many-to-many matching game where transfers are endogenous, but not in the data. The maximum score estimator maximizes the number of inequalities implied by pairwise stability that hold true. This approach breaks the computational curse of dimensionality because not all inequalities need to be included for the estimator to be consistent.

Logit Methods

Dagsvik (2000) and Choo and Siow (2006) study games with transfers, and assume that the payoffs to matches have error terms that satisfy the parametric logit property. To a large degree, the logit assumption allows researchers to derive closed-form equations that allow estimation, especially for very large markets that plausibly have a continuum of agents, such as the US national marriage market.

Identification

In matching games, agents on the same side of the market are rivals to match with agents on the other side of the market. The fact that a man did not match with the most attractive woman does not mean that the man did not prefer that woman to his actual wife. The equilibrium budget set of each agent is unobserved. Thus, it is not clear what can be learned (identified) from data on who matches with whom.

Fox (2009b) studies identification in matching games with transfers and finds two sets of results. First, the relative importance of complementarities in payoffs for say schooling, compared with say wealth, is identified using data on matches but not the equilibrium transfers that are present in the model. Second, the ordering of production levels (which matches give higher payoffs) is identified using the same data.

Selection Correcting Outcome Equations

Sørensen (2007) explores the use of a matching model to parametrically selection correct an auxiliary outcome equation. The outcome, the success of an investment in his application, is not determined as part of the matching game, but the outcome is only observed in the data for realized matches. Sørensen's approach is analogous to using a single agent decision model to selection correct an outcome equation (Heckman 1979). Sørensen (2009) extends the framework of Heckman (1990) to study identification in selection models where selection is induced by a matching game.

Empirical Applications

I now catalogue some of the many empirical applications of matching games.

Marriage and Family Economics

Choo and Siow (2006) explore whether changing matching patterns in the US marriage market are due to changes in preferences or changes in the exogenous characteristics of potential spouses. They also explore the effects of the legalization of abortion. There have been a large number of marriage and family economics papers following up on the Choo and Siow framework, many by the original authors. See Siow (2008) for a complete survey of this material.

Bruze (2009) estimates a matching game where labour supply and the split of consumption between men and women are part of the pairwise stable contract terms. He explores the return to an agent for finding a higher-earning spouse in college.

Hitsch et al. (2009) use revealed preference information from an online dating site to avoid the need to use an equilibrium model to estimate preferences. They use the preference estimates to simulate a pairwise stable outcome and find it matches well with actual sorting on the site.

Industrial Organization, Corporate Finance and Marketing

Hall (1988) is an early paper that emphasizes the need for matching models to study mergers. She did not estimate a full matching game because of computational concerns.

It is common to have data on realized interfirm relationships. Sørensen (2007) studies the matching of venture capitalists to entrepreneurs with a focus on selection correcting an outcome equation where the success of an investment is regressed upon the experience of the venture capitalist. The basic approach of Sørensen allows for match-specific error terms, and he can allow for time-invariant characteristics of a venture capitalist by using fixed effects/panel data. Chen (2009) uses a similar selection-correction framework where the outcome equation of interest is the price of a bank loan. Akkus (2008) uses the selection-correction approach to regress the degree of first-day underpricing on the experience of an investment bank. Park (2008) uses a similar MCMC estimator to investigate the decision of a mutual fund manager to engage in a merger as a function of past returns.

Fox and Bajari (2009) was the first paper to estimate a many-to-one matching game where complementarities across multiple matches were allowed for. The authors look at auctions of multiple heterogeneous items, where each bidder can win multiple items. They study FCC spectrum auctions, where complementarities between the geographic territories being auctioned are estimated to be important for the efficient operation of the mobile phone industry. A key methodological challenge is showing how a potentially inefficient, dynamic Nash game could result in equilibria that satisfy pairwise stability. The estimator used is that of Fox (2009a) for matching games with transfers. Fox (2009a) studies the many-to-many matching of automotive suppliers to automotive assemblers, and measures the relative importance of specialization by suppliers in particular corporations, brands and car models. Further, Fox measures a potential benefit of suppliers matching with high-quality Asian assemblers, such as Toyota.

Levine (2008) uses the estimator of Fox (2009a) to explore the matching of biotechnology innovations to marketing firms. She explores whether the returns to scale of marketers might decrease the returns to innovators. Yang et al. (2009) use the same estimator to explore the matching of professional athletes to teams, with a focus on the potential marketing complementarities between players and teams from different-sized cities. Akkus and Hortaçsu (2007) extend the maximum score estimator to use data on equilibrium transfers. Akkus and Hortaçsu investigate geographic complementarities in the market for bank mergers after the removal of prohibitions against interstate banks. Mindruta (2009) studies the matching of university researchers and private firms.

Development, Public Finance, Labour Economics and Other

Boyd and colleagues (2003) investigate the matching of teachers to public schools, with a focus on learning how to attract qualified teachers to schools in impoverished areas.

Gordon and Knight (2009) investigate the consolidation decisions of Iowa school districts after the state passed incentives inducing such consolidation.

Ahlin (2009) uses the estimator of Fox (2009a) to study the matching of Thai villagers to other villagers in risk-sharing groups. He investigates whether villagers match by risk type or seek to diversify risk.

Fox (2008) estimates a repeated matching model for the labour market for engineers in Sweden. Each period state variables evolve, a matching model opens, prices are formed to equate supply and demand and workers choose jobs. The model is dynamic in that both firms and workers are forward looking: they consider the effect of the decision to switch today on future outcomes.

Baccara and colleagues (2009) study the matching of professors to offices, and estimate the importance of various types of professional networks in payoffs.

See Also

- ▶ Assortative Matching
- ▶ Econometrics
- ▶ Marriage Markets
- ▶ Matching

Bibliography

Ahlin, C. 2009. Matching for credit: Risk and diversification in Thai micro-borrowing groups. Michigan State University working paper, September.

Akkus, O. 2008. Endogenous matching and the determinants of underpricing in IPO markets. University of Chicago working paper, November.

Akkus, O. and Hortaçsu, A. 2007. The determinants of bank mergers: A revealed preference analysis. University of Chicago working paper, March.

Atakan, A.E. 2006. Assortative matching with explicit search costs. *Econometrica* 74: 667–81.

Baccara, M., A. Imrohoroglu, A. Wilson, and L. Yariv. 2009. A field study on matching with network externalities. New York University working paper, August.

Becker, G.S. 1973. A theory of marriage: Part I. *Journal of Political Economy* 81: 813–46.

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

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff. 2003. Analyzing determinants of the matching of public school teachers to jobs: Estimating compensating differentials in imperfect labor markets. University of Buffalo working paper.

Bresnahan, T.F., and P.C. Reiss. 1991. Empirical models of discrete games. *Journal of Econometrics* 48: 57–81.

Bruze, G. 2009. Schooling, marriage, and male and female consumption. Aarhus University working paper, March.

Chen, J. 2009. Two-sided matching and spread determinants in the loan market. University of California at Irvine working paper, May.

Choo, E., and A. Siow. 2006. Who marries whom and why. *Journal of Political Economy* 114: 175–201.

Crawford, V.P., and E.M. Knoer. 1981. Job matching with heterogeneous firms and workers. *Econometrica* 49: 437–50.

Dagsvik, J.K. 2000. Aggregation in matching markets. *International Economic Review* 41: 27–57.

Fox, J.T. 2008. An empirical, repeated matching game applied to market thickness and switching. University of Chicago working paper.

Fox, J.T. 2009a. Estimating matching games with transfers. University of Chicago working paper, August.

Fox, J.T. 2009b. Identification in matching games. University of Chicago working paper, June.

Fox, J.T. and P. Bajari. 2009. Measuring the efficiency of an FCC spectrum auction. University of Chicago working paper.

Gale, D., and L. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69.

Gordon, N., and B. Knight. 2009. A spatial merger estimator with an application to school district consolidation. *Journal of Public Economics* 93: 752–65.

Hall, B.H. 1988. Estimation of the probability of acquisition in an equilibrium setting. University of California at Berkeley working paper, August.

Hatfield, J.W., and P.R. Milgrom. 2005. Matching with contracts. *American Economic Review* 95: 913–35.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–62.

Heckman, J.J. 1990. Varieties of selection bias. *American Economic Review* 80: 313–18.

Hitsch, G., A. Hortaçsu, and D. Ariely. 2009. Matching and sorting in online dating markets. *American Economic Review.*

Kelso, A.S., and V.P. Crawford. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica* 50: 1483–1504.

Koopmans, T.C., and M. Beckmann. 1957. Assignment problems and the location of economic activities. *Econometrica* 25: 53–76.

Levine, A.A. 2008. Licensing and scale economies in the biotechnology pharmaceutical industry. Harvard University working paper, June.

Mindruta, D. 2009. Value creation in university-firm research collaborations: A matching approach. HEC Paris working paper, February.
Park, M. 2008. An empirical two-sided matching model of acquisitions: Understanding merger incentives and outcomes in the mutual fund industry. University of Minnesota working paper, September.
Reiss, P.C., and F.A. Wolak. 2007. Structural econometric modeling: Rationales and examples from industrial organization. In *Handbook of Econometrics, Vol. 6A*. Elsevier.
Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82: 34.
Roth, A.E., and M.A.O. Sotomayor. 1990. Two-sided matching number 18. Econometric Society Monograph, Econometric Society.
Shapley, L.S., and M. Shubik. 1972. The assignment game I: The core. *International Journal of Game Theory* 1: 111–30.
Shimer, R., and L. Smith. 2000. Assortative matching and search. *Econometrica* 68: 343–69.
Siow, A. 2008. How does the marriage market clear? An empirical framework. University of Toronto working paper, July.
Sørensen, M. 2007. How smart is smart money? A two-sided matching model of venture capital. *Journal of Finance* 62: 2725–62.
Sørensen, M. 2009. Identification of multi-index selection models. Columbia University working paper.
Yang, Y., M. Shi, and A. Goldfarb. 2009. Estimating the value of brand alliances in professional team sports. *Marketing Science*.

Material Balances

Gregory Grossman

Keywords
Command economy; Material balances; Planning

JEL Classifications
P2

A material balance is a simple planning device developed (if not originated) early in Soviet planning for the purpose of equating prospective availabilities of a given good and its prospective requirements over the plan period (or at some target date in case of a stock). It occupies a central role in Soviet-type planning. The phrase, a literal rendering of the Russian *material'nyi balans*, is somewhat inexact and possibly confusing inasmuch as each of the two words has a variety of meanings in English. A more exact term would be 'sources-and-uses account' for a flow or 'balance sheet' for a stock. As such, material balances have counterparts in planning and management the world over.

In Soviet-type planning, a material balance is typically constructed *ex ante*. It can pertain to any good or resource requiring planners' attention or administrative disposition; thus, 'balance' is drawn up not only for material products, but also for labour, capacity, foreign exchange, and so on. While it can be drawn up at any level of the hierarchy of a command economy and by any relevant organizational entity, these alternatives carry important economic, bureaucratic and even political implications in a Soviet-type economy. 'In the course of preparing the annual plan ... the USSR State Planning Commission draws up [some] 2,000 single-product balances, the State Commission for Supply – up to 15,000, and the ministries – up to 50,000' (*EKO*, August, 1983, p. 26). Though there may be some duplication in terms of goods between these figures, they nonetheless do suggest the magnitude of the annual task, especially if one bears in mind the interconnections.

In Soviet-type practice a material balance not only has the passive purpose of checking requirements against availabilities, but forms the operational basis for specific production or import directives to designated organizations and firms, and for specific acquisition permits to designated users of the good. Note that nearly all producer goods are administratively allocated (rationed) to users Table 1.

A material balance may take the following form (adapted from Levine, 1959):

Two kinds of questions arise: (*a*) operational – how is the balance initially compiled and 'balanced', and later adjusted for outside effects (from other balances) and the extent to which successive iterations are required to converge?

Material Balances, Table 1

Material balance for good X for (year)	
Sources	Uses (distribution)
1. Current production – by major producing organizations, firms	1. For production – by organizations, firms
2. Imports	2. For construction – by organizations, firms
3. Other sources	3. For household sector('market fund')
4. Beginning-year stocks – by organizations	4. For export
5. Total sources	5. To central reserve stocks
	6. End-year stocks at suppliers – by organizations, firms
	7. Total uses (distribution)

and (*b*) policy – the bounds and degree of aggregation of a 'good', the organizational locus and level of compilation, and so on?

Little is known about the initial compilation. There must be serious problems of the requisite detailed information in the case of many goods, given that the preparation of the annual plan extends over most of the pre-plan year (and often into the plan year). Thus, the database may anticipate the plan year by one-and-a-half to two years whose projection is obviously subject to uncertainty. A common problem is the uncertainty of going-on-stream of capacity under construction. Also, the data may not be very accurate to start with, given the cat-and-mouse game that firms and other subordinates play with their superiors. What is more, thousands of balances are being drawn up simultaneously, often by different organizations or subdivisions, with the obvious difficulty of mutual coordination.

The 'balancer' must take into account – in addition to technical parameters – political and other high-level decisions, existing economic programmes, bureaucratic politics, and the usual pressure to squeeze more out of the economy's resources. Corruption is not unknown. The work is largely done manually and inevitably to some extent subjectively. While computers are beginning to be used, the input–output technique – which in principle is eminently suitable for the purpose – seems to be applied for the grosser computations and checks, not for the drawing up of operational, short-term material balances. The main reasons are that the sectors in even the largest matrices are too aggregative for the material balances, and the data underlying the technical coefficients are not current enough.

Among the balancer's technical parameters, pride of place is occupied by the 'norm' – a disaggregated input–output ratio, which assists the compiler in filling in parts of both sides of the account. Much effort goes into computation of the norms, given their crucial role in the preparation of plans and the issuing of specific assignments. They are supposed to be 'scientific', that is, representing the best applicable engineering practice (note: for technical rather than economic efficiency), but given their enormous number and informational problems, this remains an ideal. In the event, the balancer must employ short-cuts and resort to optimistic assumptions in order to achieve equality of requirements and availabilities while under pressure to deliver high ('taut') production targets. A common and much criticized short-cuts is simply to raise output targets of all producers by a uniform percentage, with corresponding adjustments of the norms.

The weakest link in the material balance method is coordination among the many balances to achieve a reasonably internally consistent plan for the whole economy or a sector thereof. (Montias, 1959, discusses this at length.) Even if the implicit inter-industry matrix is close to triangular, every iteration is a major undertaking under the actual conditions. Aggregating the goods would simplify the iteration process, but would not suit well the demands posed by detailed production assignments and allocation orders. So would the holding of ample reserve stocks, which are not always there or accessible. In fact,

adjustments and corrections tend ordinarily to be carried to only a few adjoining balances.

The overall annual plan that emerges is typically of low internal consistency (not to say, economic efficiency), causing considerable difficulties to those charged with its implementation and necessitating continual further correction and adjustment during the plan year, with the same effect.

See Also

- ▶ Command Economy
- ▶ Planning
- ▶ Socialist Economies

Bibliography

Levine, H.S. 1959. The centralized planning of supply in Soviet industry. In *Comparison of the United States and soviet economies I*, ed. U.S. Congress, Joint Economic Committee. Washington, DC.

Montias, J.M. 1959. Planning with material balances in Soviet-type economies. *American Economic Review* 49: 963–985.

Mathematical Economics

Gerard Debreu

Abstract

A summary of the emergence and triumph of mathematical economics. The modern phase was deeply influenced by John von Neumann's article of 1928 on games and his paper of 1937 on economic growth. His 1944 *Theory of Games and Economic Behavior*, coauthored by Oskar Morgenstern, went beyond differential calculus and linear algebra and paved the way for the axiomatization of economic theory. This has enabled researchers to use precisely stated and flawlessly proved results, in the quest for the most direct link between the assumptions and the conclusions of a theorem. Economic theory is fated for a long mathematical future.

Keywords

Activity analysis; Allais, M.; Samuelson, P.; Anderson, R.; Arrow, K.; Asymptotic equality; Aumann, R.; Axiomatized theory; Brouwer's fixed point theorem; Brown, D.; Commodity space; Competitive equilibrium; Contract curve; Convexity; Cores; Cournot, A.; Debreu, G.; Differential calculus; Econometrics; Edgeworth, F.; Existence of general equilibrium; First theorem of welfare economics; Fixed point theorems; General equilibrium; Geometry; Hicks, J.; Hyperplanes; Implicit prices; Kakutani's fixed point theorem; Koopmans, T.; Lebesgue measure; Linear programming; Lyapunov's theorem; Mathematical economics; McKenzie, L.; Nash, J.; Measure theory; Minimax theorem; Morgenstern, O.; Nonstandard analysis; Pareto, V.; Price space; Real vector space; Robinson, A.; Sard's theorem; Scarf, H.; Set of negligible agents; Shubik, M.; Sonnenschein, H.; Uncertainty; Uniqueness of equilibrium; Vind, K.; von Neumann, J.; von Neumann's lemma; Wald, A.; Walras, L.; Walras's Law

JEL Classifications

B4

I. The steady course on which mathematical economics has held for the past four decades sharply contrasts with its progress during the preceding century, which was marked by several major scientific accidents. One of them occurred in 1838, at the beginning of that period, with the publication of Augustin Cournot's *Recherches sur les principes mathématiques de la théorie des richesses.* By its mathematical form and by its economic content, his book stands in splendid isolation in time; and in explaining its data historians of economic analysis in the first half of the 19th century must use a wide confidence interval.

The University of Lausanne was responsible for two other of those accidents. When Léon Walras delivered his first professorial lecture

there on 16 December 1870, he had held no previous academic appointment; he had published a novel and a short story but he had not contributed to economic theory before 1870; and he was exactly 36. The risk that his university took was vindicated by the appearance of the *Eléments d'économie politique pure* in 1874–7. For Vilfredo Pareto, who succeeded Walras in his chair in 1893, it was also a first academic appointment; he had not contributed to economic theory before 1892; and he was 45. This second gamble of the University of Lausanne paid off when Pareto's *Cours d'économie politique* appeared in 1896–97, followed by his *Manuel d'économie politique* in 1909, and by the article 'Economie mathématique' in 1911.

In the contemporary period of development of mathematical economics, profoundly influenced by John von Neumann, his article of 1928 on games and his paper of 1937 on economic growth also stand out as major accidents, even in a career with so many facets.

The preceding local views would yield a distorted historical perception, however, if they were not complemented by a global view which sees in the development of mathematical economics a powerful, irresistible current of thought. Deductive reasoning about social phenomena invited the use of mathematics from the first. Among the social sciences, economics was in a privileged position to respond to that invitation, for two of its central concepts, commodity and price, are quantified in a unique manner, as soon as units of measurement are chosen. Thus for an economy with a finite number of commodities, the action of an economic agent is described by listing his input, or his output, of each commodity. Once a sign convention distinguishing inputs from outputs is made, the action of an agent is represented by a point in the commodity space, a finite-dimensional real vector space. Similarly the prices in the economy are represented by a point in the price space, the real vector space dual of the commodity space. The rich mathematical structure of those two spaces provides an ideal basis for the development of a large part of economic theory.

Finite dimensional commodity and price spaces can be, and usually are, identified and treated as a Euclidean space. The stage is thus set for geometric intuition to take a lead role in economic analysis. That role is manifest in the figures that abound in the economics literature, and some of the great theorists have substituted virtuosity in reasoning on diagrams for the use of mathematical form. As for mathematical economists, geometric insight into the commodity-price space has often provided the key to the solution of problems in economic theory.

The differential calculus and linear algebra were applied to that space at first as a matter of course. By the time John Hicks's *Value and Capital* appeared in 1939, Maurice Allais' *A la recherche d'une discipline économique* in 1943, and Paul Samuelson's *Foundations of Economic Analysis* in 1947, they had both served economic theory well. They would serve it well again, but the publication of the *Theory of Games and Economic Behavior* in 1944 signalled that action was also going to take new directions. In mathematical form, the book of von Neumann and Oskar Morgenstern set a new level of logical rigour for economic reasoning, and it introduced convex analysis in economic theory by its elementary proof of the MiniMax theorem. In the next few years convexity became one of the central mathematical concepts, first in activity analysis and in linear programming, as the *Activity Analysis of Production and Allocation* edited by Tjalling Koopmans attested in 1951, and then in the mainstream of economic theory. In consumption theory as in production theory, in welfare economics as in efficiency analysis, in theory of general economic equilibrium and in the theory of the core, the picture of a convex set supported by a hyperplane kept reappearing, and the supporting hyperplane theorem supplied a standard technique for obtaining implicit prices. The applications of that theorem to economics were a ready consequence of the real vector space structure of the commodity space; yet they were made more than thirty years after Minkowski proved it in 1911.

Algebraic topology entered economic theory in 1937, when von Neumann generalized Brouwer's fixed point theorem in a lemma devised to prove the existence of an optimal growth path in his model. The lag from Brouwer's

result of 1911 to its first economic application was shorter than for Minkowski's result. It should, however, have been significantly longer, for von Neumann's lemma was far too powerful a tool for his proof of existence. Several authors later obtained more elementary demonstrations, and David Gale in particular based his in 1956 on the supporting hyperplane theorem. Thus von Neumann's lemma, reformulated in 1941 as Kakutani's fixed point theorem, was an accident within an accidental paper. But in a global historical view, the perfect fit between the mathematical concept of a fixed point and the social science concept of an equilibrium stands out. A state of a social system is described by listing an action for each one of its agents. Considering such a state, each agent reacts by selecting the action that is optimal for him given the actions of all the others. Listing those reactions yields a new state, and thereby a transformation of the set of states of the social system into itself is defined. A state of the system is an equilibrium if, and only if, it is a fixed point of that transformation. More generally, if the optimal reactions of the agents to a given state are not uniquely determined, one is led to associate a set of new states, instead of a single state, with every state of the system. A point-to-set transformation of the set of states of the social system into itself is thereby defined; and a state of the system is an equilibrium if, and only if, it is a fixed point of that transformation. In this view, fixed point theorems were slated for the prominent part they played in game theory and in the theory of general economic equilibrium after John Nash's one-page note of 1950.

A perfect fit of mathematical form to economic content was also found when the traditional concept of a set of negligible agents was formulated exactly. In 1881, in *Mathematical Psychics,* Francis Edgeworth had studied in his box the asymptotic equality of the 'contract curve' of an economy and of its set of competitive allocations. Basic to his proof of convergence is the fact that in his limiting process every agent tends to become negligible. A long period of neglect of his contribution ended in 1959, when Martin Shubik brought out the connection between the contract curve and the game theoretic concept of the core. After the second impulse given in 1962 by Herbert Scarf's first extension of Edgeworth's result, a new phase of development of the economic theory of the core was under way; and in 1964 Robert Aumann formalized the concept of a set of negligible agents as the unit interval of the real line with its Lebesgue measure. The power of that formulation was demonstrated as Aumann proved that in an exchange economy with that set of agents, the core and the set of competitive allocations coincide. Karl Vind then gave, also in 1964, a different formulation of this remarkable result in the context of a measure space of agents without atoms, and showed that it is a direct consequence of Lyapunov's theorem of 1940 on the range of an atomless vector measure. The convexity of that range explains the convexing effect of large economies. In the important case of a set of negligible agents, it justifies the convexity assumption on aggregate sets to which economic theory frequently appeals. A privileged place was clearly marked for measure theory in mathematical economics.

An alternative formulation of the concept of a set of negligible agents was proposed by Donald Brown and Abraham Robinson in 1972 in terms of Non-standard Analysis, created by Robinson in the early 1960s. Innovations in the mathematical tools of economic theory had not always been immediately and universally adopted in the past. In this case the lag from mathematical discovery to economic application was exceptionally short, and Non-standard Analysis had not been widely accepted by mathematicians themselves. Predictably the intrusion of this strange, sophisticated new tool in economic theory was greeted mostly with indifference or with scepticism. Yet it led to the form given by Robert Anderson to inequalities on the deviation of core allocations from competitive allocations, which are central to the theory of the core. In the article published by Anderson in 1978 those inequalities are stated and proved in an elementary manner, but their expression was found by means of Non-standard Analysis.

The differential calculus, which had been used earlier on too broad a spectrum of economic problems, turned out in the 1970s to supply the proper mathematical machinery for the study of the set of competitive equilibria of an economy. A partial

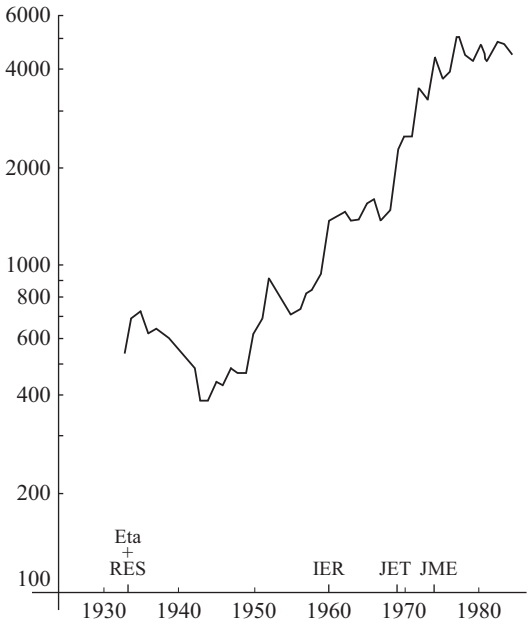
explanation of the observed state of an economic system had been provided by proofs of existence of equilibrium based on fixed point theorems. A more complete explanation would have followed from persuasive assumptions on a mathematical model of the economy ensuring uniqueness of equilibrium. Unfortunately the assumptions proposed to that end were excessively stringent, and the requirement of global uniqueness had to be relaxed to that of local uniqueness. Even then an economy composed of agents on their best mathematical behaviour (for instance each having a concave utility function and a demand function both indefinitely differentiable) may be ill-behaved and fail to have locally unique equilibria. If one considers the question from the generic viewpoint, however, one sees that the set of those ill-behaved economies is negligible. This time the ideal mathematical tool for the proof of that assertion is Sard's theorem of 1942 on the set of critical values of a differentiable function. By providing appropriate techniques for the study of the set of equilibria, differential topology and global analysis came to occupy in mathematical economics a place that seemed to have been long reserved for them.

As new fields of mathematics were introduced into economic theory and solved some of its fundamental problems, a growth-generating cycle operated. The mathematical interest of the questions raised by economic theory attracted mathematicians who in turn made the subject mathematically more interesting. The resulting expansion of mathematical economics was unexpectedly rapid. Attempting to quantify it, one can use as an index the total number of pages published yearly by the five main periodicals in the field: *Econometrica* and the *Review of Economic Studies* (which both started publishing in 1933), the *International Economic Review* (1960), the *Journal of Economic Theory* (1969), and the *Journal of Mathematical Economics* (1974). The graph of that index is eloquent. It shows a first phase of decline to 1943, followed by a 33-year period of exuberant, nearly exponential growth. The annual rate of increase that would carry the index exponentially from its 1944 level to its 1977 level is 8.2 per cent, a rate that implies doubling in slightly less than nine years and that cannot easily be sustained. The years 1977–84 have indeed marked a pause that will soon resemble a stagnation phase if it persists. Among its imperfections the index gives equal weights to *Econometrica*, the *Review of Economic Studies,* and the *International Economic Review,* all of which publish articles on econometrics as well as on mathematical economics, and to the *Journal of Economic Theory* and the *Journal of Mathematical Economics*, which do not. But given lower relative weights to the first three yields even higher annual rates of exponential growth of the index for the period 1944–77.

The sweeping movement that took place from 1944 to 1977 suggests an inevitable phase in the evolution of mathematical economics. The graph illustrating that phase hints at the deep transformation of departments of economics during those 33 years. It also hints at the proliferation of discussion papers and at the metamorphosis of professional journals like the *American Economic Review*, which was almost pure of mathematical symbols in 1933 but had lost its innocence by the late 1950s (Fig. 1).

II. As a formal model of an economy acquires a mathematical life of its own, it becomes the object of an inexorable process in which rigour, generality and simplicity are relentlessly pursued.

Before 1944, articles on economic theory only exceptionally met the standards of rigour common in mathematical periodicals. But several of the exceptions were outstanding, among them the two papers of von Neumann of 1928 and of 1937, and the three papers of Abraham Wald of 1935–6 on the existence of a general economic equilibrium. In 1944 the *Theory of Games and Economic Behavior* gained full rights for uncompromising rigour in economic theory and prepared the way for its axiomatization. An axiomatized theory first selects its primitive concepts and represents each one of them by a mathematical object. For instance the consumption of a consumer, his set of possible consumptions and his preferences are represented respectively by a point in the commodity space, a subset of the commodity space and a binary relation in that subset. Next, assumptions on the objects representing the primitive concepts are specified,



Mathematical Economics, Fig. 1 Number of pages published yearly by the leading journals in mathematical economics (Econometrica (abbr. Eta), *Review of Economic Studies*, (For the first 29 years the *Review of Economic Studies* was published on an academic rather than on a calendar year basis. As a result, only one issue appeared in 1933, compared with three in 1934; hence the spurious initial increase in the graph.) *International Economic Review, Journal of Economic Theory, Journal of Mathematical Economics*)

and consequences are mathematically derived from them. The economic interpretation of the theorems so obtained is the last step of the analysis. According to this schema, an axiomatized theory has a mathematical form that is completely separated from its economic content. If one removes the economic interpretation of the primitive concepts, of the assumptions and of the conclusions of the model, its bare mathematical structure must still stand. This severe test is passed only by a small minority of the papers on economic theory published by *Econometrica* and by the *Review of Economic Studies* during their first decade.

The divorce of form and content immediately yields a new theory whenever a novel interpretation of a primitive concept is discovered. A textbook illustration of this application of the axiomatic method occurred in the economic theory of uncertainty. The traditional characteristics of a commodity were its physical description, its date, and its location when in 1953 Kenneth Arrow proposed adding the state of the world in which it will be available. This reinterpretation of the concept of a commodity led, without any formal change in the model developed for the case of certainty, to a theory of uncertainty which eventually gained broad acceptance, notably among finance theorists.

The pursuit of logical rigour also contributed powerfully to the rapid expansion of mathematical economics after World War II. It made it possible for research workers to use the precisely stated and flawlessly proved results that appeared in the literature without scrutinizing their statements and their proofs in every detail. Another cumulative process could thus gather great momentum.

The exact formulation of assumptions and of conclusions turned out, moreover, to be an effective safeguard against the ever-present temptation to apply an economic theory beyond its domain of validity. And by the exactness of that formulation, economic analysis was sometimes brought closer to its ideology-free ideal. The case of the two main theorems of welfare economics is symptomatic. They respectively give conditions under which an equilibrium relative to a price system is a Pareto optimum, and under which the converse holds. Foes of state intervention read in those two theorems a mathematical demonstration of the unqualified superiority of market economies, while advocates of state intervention welcome the same theorems because the explicitness of their assumptions emphasizes discrepancies between the theoretic model and the economies that they observe.

Still another consequence of the axiomatization of economic theory has been a greater clarity of expression, one of the most significant gains that it has achieved. To that effect, axiomatization does more than making assumptions and conclusions explicit and exposing the deductions linking them. The very definition of an economic concept is usually marred by a substantial margin of ambiguity. An axiomatized theory substitutes for that ambiguous concept a mathematical object that is subjected to definite rules of reasoning. Thus an

axiomatic theorist succeeds in communicating the meaning he intends to give to a primitive concept because of the completely specified formal context in which he operates. The more developed this context is, the richer it is in theorems, and in other primitive concepts, the smaller will be the margin of ambiguity in the intended interpretation.

Although an axiomatic theory may flaunt the separation of its mathematical form and its economic content in print, their interaction is sometimes close in the discovery and elaboration phases. As an instance, consider the characterization of aggregate excess demand functions in an l-commodity exchange economy. Such a function maps a positive price vector into an aggregate excess demand vector, and Walras' Law says that those two vectors are orthogonal in the Euclidean commodity-price space. That function is also homogeneous of degree zero. For a mathematician, these are compelling reasons for normalizing the price vector so that it belongs to the unit sphere. Then aggregate excess demand can be represented by a vector tangent to the sphere at the price vector with which it is associated. In other words, the aggregate excess demand function is a vector field on the positive unit sphere. Hugo Sonnenschein conjectured in 1973 that any continuous function satisfying Walras' Law is the aggregate excess demand function of a finite exchange economy. A proof of that conjecture (Debreu 1974) was suggested by the preceding geometric interpretation since any vector field on the positive unit sphere can be written as a sum of l elementary vector fields, each one obtained by projecting a positive vector on one of the l coordinate axes into the tangent hyperplane. There only remains to note that every continuous elementary vector field is the excess demand function of a mathematically well-behaved consumer. Mathematical form and economic content alternately took the lead in the development of this proof.

The pursuit of generality in a formalized theory is no less imperative than the pursuit of rigour, and the mathematician's compulsive search for ever weaker assumptions is reinforced by the economist's awareness of the limitations of his postulates. It has, for example, expurgated superfluous differentiability assumptions from economic theory, and prompted its extension to general commodity spaces.

Akin in motivation, execution and consequences is the pursuit of simplicity. One of its expressions is the quest for the most direct link between the assumptions and the conclusions of a theorem. Strongly motivated by aesthetic appeal, this quest is responsible for more transparent proofs in which logical flaws cannot remain hidden, and which are more easily communicated. In extreme cases the proof of an economic proposition becomes so simple that it can dispense with mathematical symbols. The first main theorem of welfare economics, according to which an equilibrium relative to a price system is a Pareto optimum, is such a case.

In the demonstration, we study an economy consisting of a set of agents who have collectively at their disposal positive amounts of a certain number of commodities and who want to allocate these total resources among themselves. By the consumption of an agent, we mean a list of the amounts of each commodity that he consumes. And by an allocation, we mean a specification of the consumption of each agent such that the sum of all those individual consumptions equals the total resources. Following Pareto, we compare two allocations according to a unanimity principle. We say that the second allocation is collectively preferred to the first allocation if every agent prefers the consumption that he receives in the second to the consumption that he receives in the first. According to this definition, an allocation is optimal if no other allocation is collectively preferred to it. Now imagine that the agents use a price system, and consider a certain allocation. We say that each agent is in equilibrium relative to the given price system if he cannot satisfy his preferences better than he does with his allotted consumption unless he spends more than he does for that consumption. We claim that an allocation in which every agent is in equilibrium relative to a price system is optimal. Suppose, by contradiction, that there is a second allocation collectively preferred to the first. Then every agent prefers his consumption in the second allocation to his consumption in the first. Therefore the consumption of every agent in the second allocation is more

expensive than his consumption in the first. Consequently the total consumption of all the agents in the second allocation is more expensive than their total consumption in the first. For both allocations, however, the total consumption equals the total resources at the disposal of the economy. Thus we asserted that the value of the total resources relative to the price system is greater than itself. A contradiction has been obtained, and the claim that the first allocation is optimal has been established.

This result, which provides an essential insight into the role of prices in an economy and which requires no assumption within the model, is remarkable in another way. The two concepts that it relates might have been isolated, and its symbol-free proof might have been given early in the history of economic theory and without any help from mathematics. In fact that demonstration is a late by-product of the development of the mathematical theory of welfare economics. But to economists who have even a casual acquaintance with mathematical symbols, the previous exercise is not more than an artificial *tour de force* that has lost the incisive conciseness of a proof imposing no bar against the use of mathematics. That conciseness is one of the most highly prized aspects of the simplicity of expression of a mathematized theory.

In close relationship with its axiomatization, economic theory became concerned with more fundamental questions and also more abstract. The problem of existence of a general economic equilibrium is representative of those trends. The model proposed by Walras in 1874–7 sought to explain the observed state of an economy as an equilibrium resulting from the interaction of a large number of small agents through markets for commodities. Over the century that followed its publication, that model came to be a common intellectual framework for many economists, theorists as well as practitioners. This eventually made it compelling for mathematical economists to specify assumptions that guarantee the existence of the central concept of Walrasian theory. Only through such a specification, in particular, could the explanatory power of the model be fully appraised. The early proofs of existence of Wald in 1935–6 were followed by a pause of nearly two decades, and then by the contemporary phase of development beginning in 1954 with the articles of Arrow and Debreu, and of Lionel McKenzie.

In the reformulation that the theory of general economic equilibrium underwent, it reached a higher level of abstraction. From that new viewpoint a deeper understanding both of the mathematical form and of the economic content of the model was gained. Its role as a benchmark was also perceived more clearly, a role which prompted extensions to incomplete markets for contingent commodities, externalities, indivisibilities, increasing returns, public goods, temporary equilibrium, ...

In an unanticipated, yet not unprecedented, way greater abstraction brought Walrasian theory closer to concrete applications. When different areas of the field of computable general equilibrium were opened to research at the University of Oslo, at the Cowles Foundation, and at the World Bank, the algorithms of Scarf included in their lineage proofs of existence of a general economic equilibrium by means of fixed point theorems. This article has credited the mathematical form of theoretic models with many assets. Their sum is so large as to turn occasionally into a liability, as the seductiveness of that form becomes almost irresistible. In its pursuit, research may be tempted to forget economic content and to shun economic problems that are not readily amenable to mathematization. No attempt will be made here, however, to draw a balance sheet, to the debit side of which justice would not be done. Economic theory is fated for a long mathematical future, and in other editions of Palgrave authors will have the opportunity, and possibly the inclination, to choose as a theme 'Mathematical Form vs. Economic Content'.

First published in Econometrica, November 1986, with revisions.

See Also

- ▶ Computation of General Equilibria
- ▶ Cores
- ▶ Existence of General Equilibrium
- ▶ Game Theory

▶ Regular Economies
▶ Uncertainty and General Equilibrium

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique*. Paris: Imprimerie Nationale.
Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46: 1483–1487.
Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability,* ed. J. Neyman, 507–532. Berkeley: University of California Press.
Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. In *Econométrie*, 41–48. Paris: Centre National de la Recherche Scientifique.
Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.
Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.
Arrow, K.J., and M.D. Intriligator, eds. 1981–5. *Handbook of mathematical economics,* vols. 1, 2 and 3. Amsterdam: North-Holland Publishing Company.
Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.
Balasko, Y. 1986. *Foundations of the theory of general equilibrium*. New York: Academic.
Brouwer, L.E.J. 1912. Über Abbildungen von Mannigfaltigkeiten. *Mathematische Annalen* 71: 97–115.
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 USA* 69: 1258–1260.
Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: L. Hachette.
Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.
Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences of the USA* 38: 886–893.
Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New York: Wiley.
Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.
Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–21.
Debreu, G. 1977. The axiomatization of economic theory. Unpublished lecture given at the University of Bonn, 22 April.
Debreu, G. 1982. Existence of competitive equilibrium. Chapter 15 in Arrow and Intriligator (1981–5).
Dierker, E. 1974. *Topological methods in Walrasian economics*. Berlin: Springer.
Dierker, E. 1975. Gains and losses at core allocations. *Journal of Mathematical Economics* 2: 119–128.
Dierker, E. 1982. Regular economies. Chapter 17 in Arrow and Intriligator (1981–5).
Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.
Gale, D. 1956. The closed linear model of production. In *Linear inequalities and related systems*, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.
Hildenbrand, W. 1982. Core of an economy. Chapter 18 in Arrow and Intriligator (1981–5).
Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8: 457–459.
Koopmans, T.C., ed. 1951. *Activity analysis of production and allocation*. New York: Wiley.
Lyapunov, A.A. 1940. Sur les fonctions-vecteurs complètement additives. *Izvestiia Akademii Nauk SSSR* 4: 465–478.
Mantel, R. 1974. On the characterization of aggregate excess demand. *Journal of Economic Theory* 7: 348–353.
Mas-Colell, A. 1985. *The theory of general economic equilibrium: A differentiable approach*. Cambridge: Cambridge University Press.
McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.
Minkowski, H. 1911. Theorie der konvexen Körper. In *Gesammelte Abhandlungen*, vol. 3, 131–229. Leipzig/Berlin: Teubner.
Nash, J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the USA* 36: 48–49.
Pareto, V. 1896–7. *Cours d'économie politique*. Lausanne: Rouge.
Pareto, V. 1909. *Manuel d'économie politique*. Paris: Giard.
Pareto, V. 1911. Economie mathématique. In *Encyclopédie des sciences mathématiques*, vol. I(4), 591–640. Paris: Gauthier-Villars.
Robinson, A. 1966. *Non-standard analysis*. Amsterdam: North-Holland.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Sard, A. 1942. The measure of the critical points of differentiable maps. *Bulletin of the American Mathematical Society* 48: 883–890.
Scarf, H. 1962. An analysis of markets with a large number of participants. In *Recent advances in game theory*, ed. H. Scarf. Princeton: Princeton University Press.
Scarf, H. 1973. (With the collaboration of T. Hansen) *The computation of economic equilibria.* New Haven: Yale University Press.

Scarf, H. 1982. The computation of equilibrium prices: an exposition. Chapter 21 in Arrow and Intriligator (1981–5).
Scarf, H.E., and J.B. Shoven. 1984. *Applied general equilibrium analysis*. Cambridge: Cambridge University Press.
Shubik, M. 1959. Edgeworth market games. In *Contributions to the theory of games, vol. 4*, Annals of Mathematical Studies 40. Princeton: Princeton University Press.
Smale, S. 1981. Global analysis and economics. Chapter 8 in Arrow and Intriligator (1981–5).
Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 345–354.
Vind, K. 1964. Edgeworth-allocations in an exchange economy with many traders. *International Economic Review* 5: 165–177.
von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.
von Neumann, J. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. In *Ergebnisse eines mathematischen Kolloquiums*, vol. 8, 73–83. Wien: Deuticke.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Wald, A. 1935. Über die eindeutige positive Lösbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen Kolloquiums* 6: 12–20.
Wald, A. 1936a. Über die Produktionsgleichungen der ökonomischen Wertlehre. In *Ergebnisse eines mathematischen Kolloquiums*, vol. 7, 1–6. Leipzig: Deuticke.
Wald, A. 1936b. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7: 637–670.
Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: L. Corbaz.

Mathematical Methods in Political Economy

F. Y. Edgeworth

Keywords

Calculus of variations; Mathematical economics; Mathematical method in political economy; Simultaneous equations; Statistics

JEL Classifications

C0

The idea of applying mathematics to human affairs may appear at first sight an absurdity worthy of Swift's Laputa. Yet there is one department of social science which by general consent has proved amenable to mathematical reasoning – statistics. The operations not only of arithmetic, but also of the higher calculus, are applicable to statistics. What has long been admitted with respect to the average results of human action has within the last half-century been claimed for the general laws of political economy. The latter, indeed, unlike the former, do not usually present numerical constants; but they possess the essential condition for the application of mathematics: constancy of *quantitative* – though not necessarily numerical – relations. Such, for example, is the character of the law of Diminishing Returns: that an increase in the capital and labour applied to land is (tends to be) attended with a less than proportionate increase in produce. The language of Functions is well adapted to express such relations. When, as in the example given, and frequently in economics (see Marshall, *Principles,* 5th edn, Preface, p. xix), the relation is between *increments* of quantities, the differential calculus is appropriate. In the simpler cases the geometrical representations of functions and their differentials may with advantage be employed.

Among the branches of the economic calculus *simultaneous equations* are conspicuous. Given several quantitative – though not in general numerical – relations between several variable quantities, the economist needs to know whether the quantities are to be regarded as *determinate*, or not. A beautiful example of numerous prices determined by numerous conditions of supply and demand is presented by Professor Marshall in his 'bird's-eye view of the problems of joint demand, composite demand, joint supply, and composite supply' (*Principles,* Mathematical Appendix, note xxi). 'However complex the problem may become, we can see that it is theoretically determinate' (ibid., cf. Preface, p. xx). When we have to do with only *two* conditions, two *curves*

may be advantageously employed instead of two *equations.*

The mathematical operations which have been mentioned, and others – in particular the integral calculus, are all contained in the calculus of *maxima* and *minima*, or, as it is called, of *variations;* which seems to comprehend all the higher problems of abstract economics. For instance, Prof. Marshall, after writing out a number of equations 'representing the causes that govern the investment of capital and effort in any undertaking', adds, 'they may all be regarded as mathematically contained in the statement that H–V [the net advantages] is to be made a maximum' (*Principles,* Mathematical Appendix, 2nd and later editions, note xiv). It was profoundly said by Malthus, 'Many of the questions both in morals and politics seem to be of the nature of the problems *de maximis et minimis* in fluxions.' The analogy between economics and mechanics in this respect is well indicated by Dr Irving Fisher in his masterly *Mathematical Investigations*.

The property of dealing with quantities not expressible in numbers, which is characteristic of mathematical economics, is not to be regarded as a degrading peculiarity. It is quite familiar and allowed in ordinary mathematics. For instance, if one side of a plane triangle is greater than another, the angle opposite the greater side is greater than the angle opposite the less side (*Euclid*, Book I). Quantitative statements almost as loose as those employed in abstract economics occur in the less perfectly conquered portions of mathematical physics, with respect to the distances of the fixed stars, for instance (see Sir Robert Ball, *Story of the Heavens*, ch. xxi); e.g. before 1853 it was only known that 'the distance of 61 Cygni could not be *more* than sixty billions of miles'. It is really less than forty billions.

The instance of astronomy suggests a secondary or indirect use of mathematical method in economics, which physical science has outgrown. As the dawn of the Newtonian, or even of the Copernican, theory put to flight the vain shadows of astrology, so the mere statement of an economic problem in a mathematical form may correct fallacies. Attention is directed to the data which would be required for a scientific solution of the problem. Variable quantities expressed in symbols are less liable to be treated as constant. This sort of advantage is obtained by formulating the relation between quantity of precious metal in circulation and the general level of prices, as Sir John Lubbock (senior) has done in his pamphlet *On currency* (anonymous, 1840). Thus the mathematical method contributes to that negative or dialectic use of theory which consists in meeting fallacious arguments on their own ground of abstract reasoning (see some remarks on this use of theory by Prof. Simon Newcomb in the June number of the *Quarterly Journal of Economics,* 1893; and compare Prof. Edgeworth, *Economic Journal,* vol. i, p. 627). The mathematical method is useful in clearing away the rubbish which obstructs the foundation of economic science, as well as in affording a plan for the more regular part of the structure.

The modest claims here made for the mathematical method of political economy may be illustrated by comparing it with the literary or classical method in the treatment of some of the higher problems of the science. The fundamental principle of supply and demand has been stated by J.S. Mill with much precision in ordinary language (*Political Economy,* book iii, ch. 2, §§ 4, 5, and, better, review of Thornton, *Dissertations*, vol. iv). But he is not very happy in indicating the distinction between a rise of price which is due to a diminution of supply – the dispositions of the buyers, the Demand Curves remaining constant – and the rise of price which is due to a displacement of the demand curve. He appears not to perceive that the position of equilibrium between supply and demand is *determinate*, even where it is not *unique* – a conception supplied by equations with multiple roots or curves intersecting in several points. The want of this conception seems to involve even Mill's treatment of the subject in obscurity (*Political Economy,* book iii, ch. 18, § 6).

The use of simultaneous equations or intersecting curves facilitates the comprehension of the 'fundamental symmetry' (Marshall) between the forces of demand and supply; the littérateurs lose themselves in wordy disputes as to which of the two factors 'regulates' or 'determines' value.

The disturbance of the conditions of supply by a tax or bounty, or other impediment or aid, gives rise to problems too complicated for the unaided intellect to deal with. Prof. Marshall, employing the mathematical theory of Consumers' rent, reaches the conclusion that it might *theoretically* be advantageous to tax commodities obeying the law of decreasing returns in order with the proceeds to give bounty to commodities following the opposite law (*Principles,* book v, ch. xiii, § 7). The want of the theory of consumer's rent renders obscure Mill's treatment of the 'gain' which a country may draw to itself by taxing exports or imports (*Political Economy,* book v, ch. 4, § 6; cf. book iii, ch. 18, § 5). This matter is much more clearly expressed by the curves of Messrs. Auspitz and Lieben (*Untersuchungen*, Article 81).

The preceding examples presuppose free competition; the following relate to monopoly. The relation between the rates and the traffic of a railway is shown with remarkable clearness by the aid of a diagram in the appendix to Prof. Hadley's *Railroad Transportation.* By means of elaborate curves Prof. Marshall shows that a government having regard to the interest of the consuming public, as well as to its revenue, may fix a much lower price than a monopolist actuated by mere self-interest. The taxation of monopolies presents problems which require the mathematical method initiated by Cournot. His reasoning convinces of error the following statement made by Mill (book v, chs 4, 6) and others: 'A tax on rare and high-priced wines will fall only on the owners of the vineyard,' for 'when the article is a strict monopoly . . . the price cannot be further raised to compensate for the tax'. Cournot obtains by mathematical reasoning the remarkable theorem that in cases where there is a joint demand for articles monopolized by different individuals, the purchaser may come off worse than if he had dealt with a single monopolist. This case is more important than at first appears (Marshall, *Principles,* 2nd edn, book v, ch. x, § 4; 5th edn, book v, ch. xi, § 7).

Under the head of monopoly may be placed the case of two individuals or corporate units dealing with each other. The indeterminateness of the bargain in this case is perhaps best contemplated by the aid of diagrams.

These examples, which might be multiplied, seem to prove the usefulness of the mathematical method. But the estimate would be imperfect without taking into account the abuses and defects to which the method is liable. One of these is common to every *organon* – especially new ones – liability to be overrated. As Prof. Marshall says, 'When the actual conditions of particular problems have not been studied, such [mathematical] knowledge is little better than a derrick for sinking oil-wells where there are no oil-bearing strata.' Again, the mathematical method is a machinery, the use of which is very liable to be overbalanced by the cost to others than the maker of acquiring it. Not only is mathematics a foreign language 'to the general'; but even to mathematicians a new notation is an unknown dialect which it may not repay to learn. As Prof. Marshall says, 'It seems doubtful whether any one spends his time well in reading lengthy translations of economic doctrines into mathematics that have not been made by himself.'

This estimate of the uses and dangers of mathematical method may be confirmed by reference to the works in the subjoined list; which does not pretend to be exhaustive.

Bibliography

Auspitz, R., and R. Lieben. 1889. *Untersuchungen über die Theorie des Preises*. Leipzig: Duncker & Humblot.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette.

Dupuit, E.T. 1844. De la mesure de l'utilité des travaux publics. *Annales des Ponts et Chaussées*, 2nd series 8: 332–375.

Dupuit, E.T. 1849. *De l'influence des Péages*. Paris: Guillaumin.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs*. 2nd ed, 1889. Berlin: Praeger.

Jevons, W.S. 1871. *Theory of political economy.* 3rd ed, 1888. London: Macmillan.

Keynes, J.N. 1891. *The scope and method of political economy*. 3rd ed., revised. London: Macmillan, 1904.

Launhardt, W. 1885. *Mathematische Begründung der Volkwirthschaftslehre*. Leipzig.

Marshall, A. 1890. *Principles of economics*. London: Macmillan, 2nd ed., 1891; 5th ed., 1907.
Pantaleoni, M. 1889. *Principii di economia pura*. Florence: G. Barbčra.
Pareto, V. 1896. *Cours d'économie politique*. Lausanne: Rouge.
Walras, L. 1874. *Eléments d'économie politique pure*. 2nd ed. Lausanne: F. Rouge, 1889.
Wicksteed, P.H. 1888. *Alphabet of economic science*. London: Macmillan.

Mathematics and Economics

E. R. Weintraub

Abstract

The interconnection of mathematics and economics reflects changes in both the mathematics and economics communities over time. The respective histories of these disciplines are intertwined, so that both changes in mathematical knowledge and changing ideas about the nature of mathematical knowledge have effected changes in the methods and concerns of economists.

Keywords

Arrow, K; Axiomatics; Cowles Commission; Debreu, G; Econometrics; Euclid; Koopmans, T; Marshall, A; Mathematical economics; Mathematics and economics; Newton, I; Whewell, W

JEL Classification

B0

Understanding the connection between mathematics and economics is not the same as understanding the nature and role of mathematical economics. 'Mathematical economics' is the employment of mathematics in economics itself. Explaining or justifying mathematical economics often involves essentialist arguments concerning the true nature of economic objects and the true nature of the economy, as well as arguments suggesting that employing mathematics is appropriate since the underlying 'economy' is quantitative in nature. Consequently, an historical discussion of mathematical economics will be a narrative of increased sophistication over time in economics, as mathematical tools, techniques and methods move into economic discourse and enrich economic analysis.

Alternatively, one can discuss the relation between mathematics and economics in terms of separate intellectual activities performed in separate intellectual communities, and in that case one will wish to look over time at the interpenetration of the ideas and practices of the two communities across their highly permeable boundaries. The history of mathematics concerns the changing body of mathematical knowledge such as new theorems proved, new research areas opened, and new techniques developed. But the history also involves changing images of mathematical knowledge: changing perspectives and understandings, for example, about the nature of mathematical objects, what constitutes a proof, what constitutes rigour, what constitutes useful versus not useful mathematics, and so forth (see Corry 1996, p. 3). Similarly the history of economics involves a history of not only the development of economic knowledge, but the development and changes in images of economic knowledge: what constitutes the economy, what constitutes a good explanation in economics, what constitutes serious empirical work in economics, what a good model is, and so on. Consequently a discussion of the interconnection of mathematics and economics requires not just attention to the interconnection of the bodies of knowledge, as is reflected in the historical discussion of mathematical economics, but a historical discussion of the interconnection of their respective images of knowledge. Put another way, a discussion of the connection of mathematics and economics must reflect economists' changing conceptions of the image of mathematical knowledge and not just their changing understandings of the body of mathematical knowledge.

This distinction between the body of knowledge and the images of knowledge provides a different perspective on the relation between

mathematics and economics. The central point for economists to understand is that there were three distinct shifts in the image of mathematics from the beginning of the nineteenth century to the end the twentieth century.

From Geometry to Mechanics

As a starting point, consider the conditions and perspectives under which mathematics was produced early in the nineteenth century. Looking closely, we see, particularly in England, the importance of both Euclid's *Elements* and Newton's *Principia*. That is, from relatively early in the nineteenth century, through the modifications of the Cambridge Tripos in 1849, and on through the middle third of the nineteenth century at Cambridge, mathematics was understood as flowing, in its purpose and nature, from both Euclid and Newton. From Euclid one understood that geometry was the paradigm of mathematics, and that it was a path to truth. Theorems were derived from assumptions called axioms, where the truth of those assumptions was self-evident from our understanding of the physical world. To learn geometry was to understand how rigorous arguments could lead to truth. One studied mathematics, specifically geometry, as an exemplar of how one deduced truths about the world, and thus mathematics was the paradigm of deductive thought and logical ratiocination. Parallel to this view of how deductive reasoning from true premises could lead to true conclusions, Newton's *Principia* (his mathematical proofs of course were all based on Euclidian geometry – even the calculus derivations were geometrical), suggested how this kind of mathematics could also open up an understanding of the physical world. Students were required to study mathematics because it provided a way of achieving truth.

This image of mathematics is at the root of Ricardo's arithmetical models, and is present in Whewell's papers (Whewell 1829, 1831, 1850) on economics using mathematics, for Whewell himself was central in reconstructing the Cambridge Tripos around Euclidean geometry and Newton's *Principia* at mid-nineteenth century. Economics was to employ a particular kind of mathematics, Euclidean geometry, to demonstrate its propositions. Just as Newton employed geometrical proofs of his propositions, so too did Marshall. It is an interesting exercise to open Alfred Marshall's *Principles* next to the Newton's *Principia* and see the physical similarity of the proofs or demonstrations of the propositions in each book. Marshall, as Second Wrangler in the Mathematical Tripos of January 1865, had had to master both Euclid and Newton.

The first change in the image of mathematics was developed from a new conception of what mathematical truth might mean. It occurred over the second third of the nineteenth century and was then well incorporated in the Continental tradition in mathematics. That is, outside Britain there was a change in the image of mathematics between the time of Whewell's defence of mathematics in the educational process, a defence based in the notion that mathematics (*vide* Euclid, Newton) was the paradigm of certain and secure knowledge (the time of Marshall's student days), and Marshall's later time as Professor of Political Economy. The emergence of non-Euclidean geometries had made Whewell's argument about axiomatics, and inevitable truth, ring hollow long before the turn of the twentieth century. In the time of the new geometries, the difficulty of linking mathematical truth to a particular (Euclidean) geometry produced a real crisis of confidence for Victorian educational practice (Richards 1988). This first crisis prepared the late Victorian mind for the new idea that mathematical rigour had to be associated with physical argumentation. And it was this new image of mathematics in science that helps us to understand the concerns of individuals like Edgeworth and Pareto.

An emergent set of themes in mathematics developed from the increased awareness of alternatives to Euclidean geometry, and the recognition that no one set of axioms could be selected for demonstrating the truth of all mathematical propositions. Thus the success of the new rational mechanics (Lagrange's programme of applying techniques of advanced calculus to the study of motions of solids and liquids) in making sense of the world of physical systems encouraged a

refinement of the truth-producing view of mathematics. That is, in the last third of the nineteenth century, in Britain as well as Italy, France and Germany, a rigorous mathematical argument began to be seen as one based on a substrate of physical reasoning. For an argument to be rigorous, and thus believable, the mathematical structure had to be founded generally on the most successful of applied mathematical practices, namely, rational mechanics. *A valid and good and useful mathematical model was a model that had physical interpretations*. The 'marginal revolution' in economics was precisely this new understanding. One sees this very clearly in Marshall, who was at the cusp of this changed image of mathematics, for his derivations were offered using Euclidian geometry, but whose mathematical arguments about equilibrium and stability are instantiations of mechanical devices like an egg in a bowl, or a pair of scissors. Put another way, through much of the nineteenth century in British mathematics, and thus to a degree among insular British economists for whom British mathematics *was* mathematics, rigour in argument was associated with geometric proofs based on assumptions, called axioms, that could be linked to constrained optimization processes associated with particular physical systems. Rational mechanics was taken as a paradigm for what economists came to call the marginal revolution, which, however, was hardly revolutionary but rather the migration of rational mechanical ideas into economic discourse (Mirowski 1989). Thus, by the last decades of the nineteenth century one finds economists employing specific mechanical models of economic behaviour. Walras, Pareto, Marshall, Edgeworth and Fisher were producing rigorous mathematical models of economic processes, where rigour was associated with a mathematics tied to physical processes.

From Mechanics to Axiomatics

But by 1900 the images of, and styles of doing, mathematics were beginning to change again in response to new challenges in mathematics and physics. In mathematics, there were problems associated with the foundations of mathematics. There were apparent inconsistencies in set theory associated with Georg Cantor's new ideas on 'infinity' (that is, transfinite cardinals, and the continuum of real numbers), and apparent inconsistencies in the foundations of arithmetic and logic, associated with work by Frege and Peano. Similarly troubling was the failure of physics, particularly rational mechanics, to solve the new problems associated with black-body radiation, quanta and relativity. If the deterministic mechanical mode of physical argumentation was to be replaced by an alternative physical theory, what constituted a rigorous mathematical argument had to be re-described. In any event, some established areas of mathematics were no longer connected to a canonical physical model (Weintraub 2002).

Consequently, around the end of the nineteenth century, just as economists had begun to understand that constructing a mathematical science required basing argumentation of the physical reasoning of rational mechanics, and the measurement of quantities to further ground those reasoning chains, the image of mathematical knowledge was again changing. Modelling the concerns of the new physics appeared to require a new mathematics, based less on deterministic dynamical systems and more on statistical argumentation, algebra, and new beliefs about appropriate axioms for logic and arithmetic.

Just as the objects of the physical world appeared changed – gone were billiard balls, newly present were quanta – the recognition that the paradoxes of set theory and logic were intertwined led mathematicians to seek new foundations for their subject. Analysis of those foundations of set theory, logic and arithmetic, and thus the foundations of sciences based on mathematics, were now to be based on axiomatic thinking. A rigorous argument was to be one built on strong foundations, and axiomatizing the structure of theories, in both physics and mathematics, was a path to the development of those theories (Hilbert 1918). Thus, following a late nineteenth century period in which mathematical rigour was to be established by basing the mathematics on physical reasoning, around 1900 – as understanding of the physical world became less

secure – *mathematical truth was to be established not relative to physical reasoning but relative to other mathematical theories and objects*. From a physical reductionism mathematics moved to a mathematical reductionism, in the guise of one or another set of ideas about formalism: problems and paradoxes and confusions in turn-of-the-century mathematics were to be resolved by a re-conceptualization of the nature of the fundamental objects of mathematics. The images of mathematical knowledge and ideas of rigour, truth, formalization and proof all changed over this period.

It took a number of decades for this new image of mathematics to become securely established in the mathematical community. From Hilbert's 1918 call for axiomatization as the road to knowledge in mathematics and science, through the interwar years, mathematicians were slow to reframe their working concerns. So too did economists' use of mathematics in the interwar period reflect the earlier perspectives of modelling economic problems as constrained optimization demonstrations imitating nineteenth century mechanics. Beginning in the 1930s, however, a group of French mathematicians, collectively called 'Nicholas Bourbaki', began rewriting mathematics from the foundationalist perspective (Weintraub and Mirowski 1994). Mathematics was conceived of, in their project, as growing organically from very basic ideas about sets, which led inexorably to the identification of a small number of 'mother structures' (algebraic, order, and topological) from which other structures, other branches of mathematics, could be derived. Rigorous mathematics was not grounded in physical models but rather in mathematics itself. Mathematics was to concern itself with analyses of mathematical structures. Over the next few decades pure mathematics, or mathematics uncontaminated by applications and disengaged from the world of applications, gained sway in the mathematics community. It was in this period that the eminent mathematician Paul Halmos (1981) famously titled an article 'Applied mathematics is bad mathematics'. In economics, this concatenation of ideas moved into mainstream theory with the work of Gerard Debreu, Kenneth Arrow and Tjalling Koopmans. The Cowles Commission, in the 1940s at the University of Chicago, became the site for production of this kind of work in mathematical economic theory, particularly general equilibrium theory.

Yet, even as a pure mathematics was taking hold in economics, the exigencies of the Second World War and economists' involvement with scientists, engineers, and other social scientists, moved mathematical economists' concerns back from axiomatization and into what would become operations research. This, of course, was not 'pure' at all, but based on concrete problems of real systems. As Amy Dahan Dalmedico, the historian of mathematics, noted:

> The second World War initiated what I shall call 'image war' or 'representation war' concerning what mathematics was about, what it dealt with, and how. Over the course of the 1950s and 1960s, this 'war' was progressively developed until the balance of power began to shift perceptibly at the end of the 1970s and during the 1980s. This 'war' was focused mainly on the cleavage between pure and applied mathematics, and on the tacit hierarchy – of concepts as much as of values – informing these categories of 'pure' and 'applied'. (Dalmedico 2001, p. 224)

Thus, Bourbakist images of mathematics were becoming dominant in economics at the same time as the major challenge to those ideas was forming outside 'pure' theory. The image of mathematics as a discipline concerned with understanding the structures of mathematical objects was indeed dominant in the 1950s and 1960 s, not only in the United States but in a number of other countries. Yet, from the Second World War on through the cold war, applied mathematics was taking root in disparately profound ways, and was attracting more and more support in the form of grants and contracts and students. New fields of statistics, computer science and operations research flourished. Consequently, economists' ideas about mathematics began to undergo changes, as usual with some time lag, mirroring the changing images of mathematics that were reshaping interests and methods in the mathematics community itself. 'While *structure* was the emblematic term of the 1960s, *model* has now taken its place. In the physical sciences,

climatology, engineering science, economics, and the social sciences, the practice of model-building has gradually dominated the terrain. It is today absolutely massive and intrinsically bound up with numerical experimentation and simulation' (Dalmedico 2001, 249).

If the important lesson from mathematics in the first third of the nineteenth century was that economics needed to become a deductive science (as geometry was), in the late nineteenth century the lesson from mathematics was that economics needed to model itself on rational mechanics. Over the first two thirds of the 20th century the lesson was that economics was to become scientific by grounding its models and theories on a modest set of axioms concerning pure economic agents' preferences and choices. But, beginning nearly at mid-century, mathematics was re-imagining itself as a discipline that historically had developed by solving real problems presented to it from other sciences. And in a similar fashion, and partially in response to that changing image of mathematical knowledge, the notion of a serious economic science, connected to data-based reasoning, was reshaping the idea of rigorous argumentation in economics. Econometrics and applied microeconomics were to form the reconstructed core of economic science much as work in algorithmics and applied mathematics were re-commanding attention in the mathematics community. 'At the Berlin International Congress of Mathematicians in August 1998, the old opposition between the pure and the applied – still widely shared in the community – has been formulated in quite different terms: "mathematicians who build models versus those who prove theorems". (Mumford 1998). But the respect enjoyed by the former is now definitely as high as that of the latter' (Dalmedico 2001, p. 249). So too in economics, as the prestige accorded 'good work' in applied economics now rivals that accorded to work in pure theory.

See Also

- ▶ Debreu, Gerard (1921–2004)
- ▶ Existence of General Equilibrium
- ▶ Fisher, Irving (1867–1947)
- ▶ Marshall, Alfred (1842–1924)
- ▶ Mathematical Methods in Political Economy
- ▶ Whewell, William (1799–1866)

Bibliography

Corry, L. 1996. *Modern Algebra and the rise of mathematical structures*. Boston: Birkhäuser.

Dalmedico, A. 2001. An image conflict in mathematics after 1945. In *Changing images in mathematics: From the French revolution to the new millennium*, ed. U. Bottazzini and A. Dalmedico. London/New York: Routledge.

Halmos, P. 1981. Applied mathematics is bad mathematics. In *Mathematics tomorrow*, ed. L. Steen. New York/Heidelberg: Springer.

Hilbert, D. 1918. Axiomatisches Denken. *Mathematische Annalen* 78: 405–415.

Mirowski, P. 1989. *More heat than light*. New York/Cambridge: Cambridge University Press.

Mumford, D. 1998. Trends in the profession of mathematics: Choosing our directions. *Berlin Intelligencer*, ICM 2–5 August 1998.

Richards, J. 1988. *Mathematical visions: The pursuit of geometry in Victorian England*. San Diego: Academic.

Weintraub, E. 2002. *How economics became a mathematical science*. Durham: Duke University Press.

Weintraub, E., and P. Mirowski. 1994. The pure and the applied: Bourbakism comes to mathematical economics. *Science in Context* 72: 245–272.

Whewell, W. 1829/1831/1850. *Mathematical exposition of some doctrines of political economy*. Reprints of economic classics. New York: Augustus M. Kelley, 1971.

M

Mathematics of Networks

M. E. J. Newman

Abstract

The patterns of interactions, both economic and otherwise, between individuals, groups or corporations form social networks whose structure can have a substantial effect on economic outcomes. The study of social networks and their implications has a long history in the social sciences and more recently in applied mathematics and related fields. This article reviews the main developments in the area

with a focus on practical applications of network mathematics.

Keywords

Bernoulli random graph; Centrality measures; Graph theory; Milgram, S.; Moreno, J.; Network formation; Networks, mathematics of; Non-Poisson degree distributions; Perron–Frobenius theorem; Small worlds; Social interactions; Social networks

JEL Classifications

D85

In much of economic theory it is assumed that economic agents interact, directly or indirectly, with all others, or at least that they have the opportunity to do so in order to achieve a desired outcome for themselves. In reality, as common sense tells us, things are quite different. Traders in a market have preferred trading partners, perhaps because of an established history of trust, or simply for convenience. Buyers and sellers have preferred suppliers and customers. Consumers have preferred brands and outlets. And most individuals limit their interactions, economic or otherwise, to a select circle of partners or acquaintances. In many cases partners are chosen not on economic grounds but for social reasons: individuals tend overwhelmingly to deal with others who revolve in the same circles as they do, socially, intellectually or culturally.

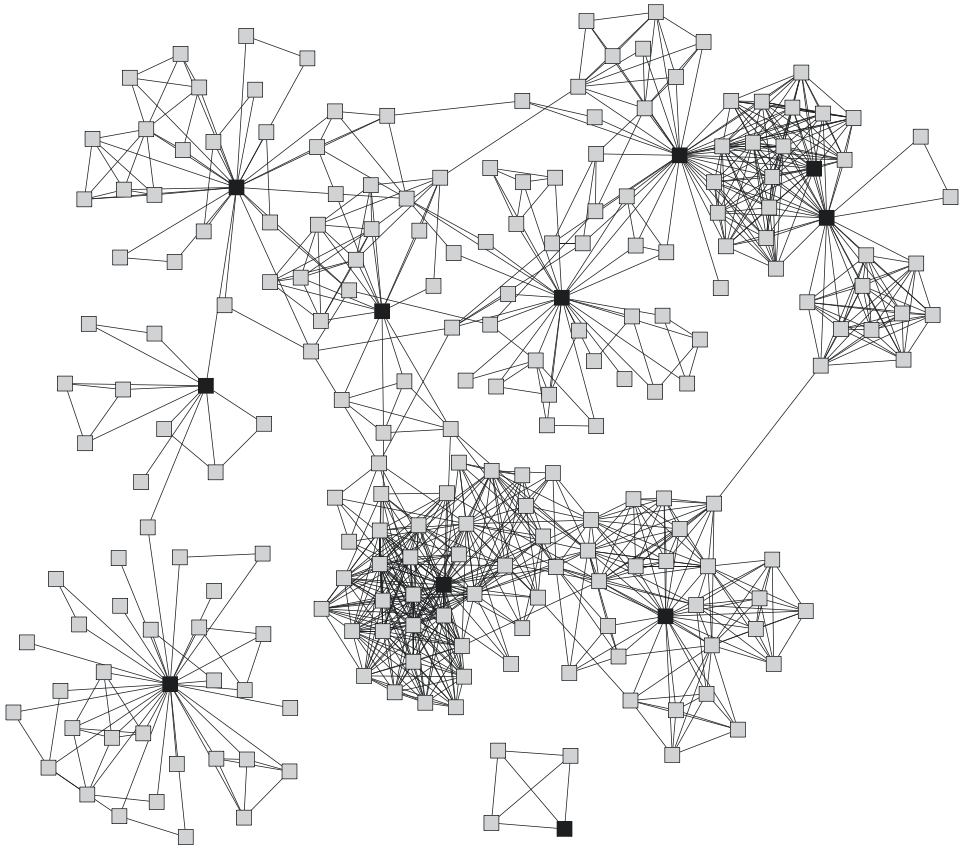
The patterns of connections between agents form a social network (Fig. 1), and it is intuitively clear that the structure of such networks must affect the pattern of economic transactions, not to mention essentially every other type of social interaction among human beings. Any theory of interaction that ignores these networks is necessarily incomplete. In the last few decades, therefore, researchers have conducted extensive investigations of networks in economics, mathematics, sociology and a number of other fields, in an effort to understand and explain network effects.

The study of social (and other) networks has three primary components. First, empirical studies of networks probe network structure using a variety of techniques such as interviews, questionnaires, direct observation of individuals, use of archival records, and specialist tools like 'snowball sampling' and 'ego-centred' studies. The goal of such studies is to create a picture of the connections between individuals, of the type shown in Fig. 1. Since there are many different kinds of possible connections between people – business relationships, personal relationships, and so forth – studies must be designed appropriately to measure the particular connections of interest to the experimenter.

Second, once one has empirical data on a network, one can answer questions about the community the network represents using mathematical or statistical analyses. This is the domain of classical social network analysis, which focuses on issues such as: who are the most central members of a network and who are the most peripheral? Which people have most influence over others? Does the community break down into smaller groups, and if so what are they? Which connections are most crucial to the functioning of a group?

And third, building on the insights obtained from observational data and its quantitative analysis, one can create models, such as mathematical models or computer models, of processes taking place in networked systems – the interactions of traders, for example, or the diffusion of information or innovations through a community. Modelling work of this type allows us to make predictions about the behaviour of a community as a function of the parameters affecting the system.

After a brief historical review, the primary purpose of this article is to describe the mathematical techniques involved in the second and third of these three components: the quantitative analysis of network data and the mathematical modelling of networked systems. Necessarily, this review is short. Much more substantial coverage can be found in the many books and review articles in the field (Wasserman and Faust 1994; Scott 2000; West 1996; Harary 1995; Ahuja et al. 1993; Dorogovtsev and Mendes 2003; Albert and Barabási 2002; Newman 2003).



Mathematics of Networks, Fig. 1 A social network of collaborative links. *Note*: The nodes (*squares*) represent people and the edges (*lines*) social ties between them

History of Social Network Analysis

The study of social networks has roots in the 19th-century beginnings of sociology, especially the 'gestalt' tradition of Koehler and others, but is widely regarded as having begun in earnest in the 1930s with the work of psychologist Jacob Moreno, a Romanian immigrant to the United States who had spent a number of years in Vienna and was influenced there by the work of Freud. Moreno advocated an approach to psychoanalysis that involved participants discussing or physically enacting issues that concerned them in front of the analyst. Another approach, which Moreno employed with schoolchildren among others, involved the analyst passively watching participants' interactions with one another and recording their nature and pattern. In the process of his studies he developed a new tool, the sociogram, which was a map of interactions between individuals drawn on paper as a set of points and lines (Moreno 1934, p. 38).

In 1933 Moreno presented some of his sociograms during a lecture at a medical conference in New York City, and the work attracted sufficient interest to be featured in the *New York Times*. In everything but name, Moreno's sociograms were what we would now call social networks, and his methods, although strange by today's standards, were the intellectual precursor of social network analysis, which is now a flourishing branch of the social sciences (Wasserman and Faust 1994). (The term 'social network' was not invented until some years later; it is usually credited to John Barnes 1954.)

Apart from a gap during the war years, social network analysis was pursued vigorously following its early popularization. Particularly well-known studies include the 'southern women' study of Davis et al. (1941), Anatol Rapoport's investigations of friendship networks among school children in the 1950s (Rapoport and Horvath 1961), Pool and Kochen's (1978) mathematical models of social networks that circulated widely in the 1950s and 1960s (although they were not published until much later), and Stanley Milgram's (1967) famous 'small world' experiments. Today, social network analysis is one of the standard quantitative tools in the social science toolbox, finding use both in academia and in the business world as a microscope with which to view the details of social interactions.

Mathematics of Networks

Turning to the mathematical methods of network analysis, which are the principal focus of this article, let us begin with some simple definitions. A network – also called a *graph* in the mathematics literature – is made up of points, usually called *nodes* or *vertices*, and lines connecting them, usually called *edges*. Mathematically, a network can be represented by a matrix called the *adjacency matrix* $\mathbf{A}$, which in the simplest case is an $n \times n$ symmetric matrix, where n is the number of vertices in the network. The adjacency matrix has elements

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The matrix is symmetric since if there is an edge between i and j then clearly there is also an edge between j and i. Thus $A_{ij} = A_{ji}$.

In some networks the edges are *weighted*, meaning that some edges represent stronger connections than others, in which case the nonzero elements of the adjacency matrix can be generalized to values other than unity to represent stronger and weaker connections. Another variant is the *directed network*, in which edges point in a particular direction between two vertices. For instance, in a network of cash sales between buyers and sellers the directions of edges might represent the direction of the flow of goods (or conversely of money) between individuals. Directed networks can be represented by an asymmetric adjacency matrix in which $A_{ij} = 1$ implies the existence (conventionally) of an edge pointing from j to i (note the direction), which will in general be independent of the existence of an edge from i to j.

Networks may also have *multiedges* (repeated edges between the same pair of vertices), *self-edges* (edges connecting a vertex to itself), *hyper-edges* (edges that connect more than two vertices together) and many other features. We here concentrate primarily on the simplest networks, having undirected, unweighted single edges between pairs of vertices.

Centrality Measures

Now let us consider the analysis of network data. We start by looking at *centrality measures*, which are some of the most fundamental and frequently used measures of network structure. Centrality measures address the question, 'Who is the most important or central person in this network?' There are many answers to this question, depending on what we mean by 'important'. Perhaps the simplest of centrality measures is *degree centrality*, also called simply *degree*. The degree of a vertex in a network is the number of edges attached to it. In mathematical terms, the degree k_i of a vertex i is

$$k_i = \sum_{j=1}^{n} A_{ij}. \tag{2}$$

Though simple, degree is often a highly effective measure of the influence or importance of a node: in many social settings people with more connections have more power.

A more sophisticated version of the same idea is the so-called *eigenvector centrality*. Where degree centrality gives a simple count of the number of connections a vertex has, eigenvector centrality acknowledges that not all connections are equal. In general, connections to people who are themselves influential will lend a person more influence than connections to less influential people. If we denote the centrality of vertex i by x_i, then we can allow for this effect by making x_i proportional to the average of the centralities of i's network neighbours:

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} A_{ij}x_j, \tag{3}$$

where λ is a constant. Defining the vector of centralities $\mathbf{x} = (x_1, x_2,..)$, we can rewrite this equation in matrix form as

$$\lambda\mathbf{x} = \mathbf{A}\cdot\mathbf{x}, \tag{4}$$

and hence we see that $\mathbf{x}$ is an eigenvector of the adjacency matrix with eigenvalue λ. On the assumption that we wish the centralities to be non-negative, it can be shown (using the Perron–Frobenius theorem) that λ must be the largest eigenvalue of the adjacency matrix and $\mathbf{x}$ the corresponding eigenvector.

The eigenvector centrality defined in this way accords each vertex a centrality that depends on both the number and the quality of its connections: having a large number of connections still counts for something, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts. Eigenvector centrality turns out to be a revealing measure in many situations. For example, a variant of eigenvector centrality is employed by the well-known Web search engine Google to rank Web pages, and works well in that context.

Two other useful centrality measures are *closeness centrality* and *betweenness centrality*. Both are based upon on the concept of network paths. A path in a network is a sequence of vertices traversed by following edges from one vertex to another across the network. A *geodesic path* is the shortest path, in terms of number of edges traversed, between a specified pair of vertices. (Geodesic paths need not be unique; two or more paths can tie for the title of shortest.) The closeness centrality of vertex i is the mean geodesic distance (that is, the mean length of a geodesic path) from vertex i to every other vertex. Closeness centrality is *lower* for vertices that are more central in the sense of having a shorter network distance on average to other vertices. (Some writers define closeness centrality to be the reciprocal of the average so that higher numbers indicate greater centrality. Also, some vertices may not be reachable from vertex i – two vertices can lie in separate 'components' of a network, with no connection between the components at all. In this case closeness as above is not well defined. The usual solution to this problem is simply to define closeness to be the average geodesic distance to all *reachable* vertices, excluding those to which no path exists.)

The betweenness centrality of vertex i is the fraction of geodesic paths between other vertices that i lies on. That is, we find the shortest path (or paths) between every pair of vertices, and ask on what fraction of those paths vertex i lies. Betweenness is a crude measure of the control i exerts over the flow of information (or any other commodity) between others. If we imagine information flowing between all pairs of individuals in the network and always taking the shortest possible path, then betweenness centrality measures the fraction of that information that will flow through i on its way to wherever it is going. In many social contexts a vertex with high betweenness will exert substantial influence by virtue not of being in the middle of the network (although it may be) but of lying 'between' other vertices in this way. It is in most cases only an approximation to assume that information flows along geodesic paths; normally it will not, and variations of betweenness centrality such as 'flow betweenness' and 'random walk betweenness' have been proposed to allow for this. In many practical cases, however, the simple (geodesic path) betweenness centrality gives quite informative answers.

Other Network Properties

The study of shortest paths on networks also leads to another interesting network concept, the *small-world effect*. It is found that in most networks the mean geodesic distance between vertex pairs is small compared with the size of the network as a whole. In a famous experiment conducted in the 1960s, the psychologist Stanley Milgram (1967) asked participants (located in the United States) to get a message to a specified target person elsewhere in the country by passing it from one acquaintance to another, stepwise through the population. Milgram's remarkable finding that the typical message passed though just six people on its journey between (roughly) randomly chosen initial and final individuals has been immortalized in popular culture in the phrase 'six degrees of separation', which was the title of a 1990 Broadway play by John Guare in which one of the characters discusses the small-world effect. Since Milgram's experiment, the small-world effect has been confirmed experimentally in many other networks, both social and nonsocial.

Other network properties that have attracted the attention of researchers in recent years include network *transitivity* or *clustering* (the tendency for triangles of connections to appear frequently in networks – in common parlance, 'the friend of my friend is also my friend'), vertex similarity (the extent to which two given vertices do or do not occupy similar positions in the network), communities or groups within networks and methods for their detection, and, crucially, the distribution of vertex degrees, a topic discussed in more detail below.

Models of Networks

Turning to models of networks and of the behaviour of networked systems, we find that perhaps the simplest useful model of a network (and one of the oldest) is the *Bernoulli random graph*, often called just the *random graph* for short (Solomonoff and Rapoport 1951; Erdős and Rényi 1960; Bollobás 2001). In this model one takes a certain number of vertices n and creates edges between them with independent probability p for each vertex pair. When p is small there are only a few edges in the network, and most vertices exist in isolation or in small groups of connected vertices. Conversely, for large p almost every possible edge is present between the $\binom{n}{2}$ possible vertex pairs, and all or almost all of the vertices join together in a single large connected group. One might imagine that for intermediate values of p the sizes of groups would just grow smoothly from small to large, but this is not the case. It is found instead that there is a *phase transition* at a special value $p = 1/n$ above which a *giant component* forms, a group of connected vertices occupying a fixed fraction of the whole network, i.e., with size varying as n. For values of p less than this, only small groups of vertices exist of a typical size that is independent of n. Many real-world networks show behaviour reminiscent of this model, with a large component of connected vertices filling a sizable fraction of the entire network, the remaining vertices falling in much smaller components that are unconnected to the rest of the network.

The random graph has a major shortcoming, however: the distribution of the degrees of the vertices is quite unlike that seen in most real-world networks. The fraction p_k of vertices in a random graph having degree k is given by the binomial distribution, which becomes Poisson in the limit of large n:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} ; \frac{z^k e^{-z}}{k!}, \qquad (5)$$

where $z = (n-1)p$ is the mean degree. Empirical observations of real networks, social and otherwise, show that most have highly non-Poisson distributions of degree, often heavily right-skewed with a fat tail of vertices having unusually high degree (Albert and Barabási 2002; Dorogovtsev and Mendes 2003). These high-degree nodes or 'hubs' in the tail can, it turns out, have a substantial effect on the behaviour of a networked system.

To allow for non-Poisson degree distributions, one can generalize the random graph, specifying a

particular, arbitrary degree distribution p_k and then forming a graph that has that distribution but is otherwise random. A simple algorithm for doing this is to choose the degrees of the n vertices from the specified distribution, draw each vertex with the appropriate number of 'stubs' of edges emerging from it, and then pick stubs in pairs uniformly at random and connect them to create complete edges. The resulting model network (or more properly the ensemble of such networks) is called the *configuration model*.

The configuration model also shows a phase transition, similar to that of the Bernoulli random graph, at which a giant component forms. To see this, consider a set of connected vertices and consider the 'boundary vertices' that are immediate neighbours of that set. Let us grow our set by adding the boundary vertices to it one by one. When we add one boundary vertex to our set the number of boundary vertices goes down by 1. However, the number of boundary vertices also increases by the number of new neighbours of the vertex added, which is one less than the degree k of that vertex. Thus the total change in the number of boundary vertices is $-1 + (k-1) = k-2$. However, the probability of a particular vertex being a boundary vertex is proportional to k, since there are k times as many edges by which a vertex of degree k could be connected to our set than there are for a vertex of degree 1. Thus the average change in the number of boundary vertices when we add one vertex to our set is a weighted average $\Sigma_i k_i(k_i-2) = \sum_j k_j = \Sigma_i k_i(k_i-2) = (nz)$, where z is again the mean degree. If this quantity is less than zero, then the number of boundary vertices dwindles as our set grows bigger and will in the end reach zero, so that the set will stop growing. Thus in this regime all connected sets of vertices are of finite size. If on the other hand this number is greater than zero, then the number of boundary vertices will grow without limit, and hence the size of our set of connected vertices is limited only by the size of the network.

Thus, a giant component exists in the network if and only if

$$\langle k^2 \rangle - 2\langle k \rangle > 0, \tag{6}$$

where $\langle k \rangle = z = n^{-1}\Sigma_i k_i$ is the mean degree and $\langle k^2 \rangle = n^{-1}\Sigma_i k_i^2$ is the mean-square degree.

The mean-square degree appears over and over in the mathematics of networks. Another context in which it appears is in the spread of information (or anything else) over a network. Taking a simple model of the spread of an idea (or a rumour or a disease), imagine that each person who has heard an idea communicates it with independent probability q to each of his or her friends. If the person's degree is k, then there are $k-1$ friends to communicate the idea to, not counting the one from whom he or she heard it in the first place, so the expected number who hear it is $q(k-1)$. Performing the weighted average over vertices again, we find that the average number of people a person passes the idea onto, also called the *basic reproductive number* R_0, is

$$R_0 = q\frac{\sum_i k_i(k_i-1)}{\sum_i k_i} = q\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}. \tag{7}$$

If R_0 is greater than 1, then the number of people hearing the idea grows as it gets passed around and it will take off exponentially. If R_0 is less than 1 then the idea will die. Again, we have a phase transition, or *tipping point*, for the spread of the idea: it spreads if and only if

$$q > \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \tag{8}$$

The simple understanding behind the appearance of the mean-square degree in this expression is the following. If a person with high degree hears this idea he or she can spread it to many others, by virtue of having many friends. However, such a person is also more likely to hear the idea in the first place because of having many friends to hear it from. Thus, the degree enters twice into the process: a person with degree 10 is $10 \times 10 = 100$ times more effective at spreading the idea than a person with degree 1.

The appearance of the mean-square degree in expressions like (6) and (8) can have substantial effects. Of particular interest are networks whose degree distributions have fat tails. It is possible for

such networks to have very large values of $\langle k^2 \rangle$– in the hundreds or thousands – so that, for example, the right-hand side of Eq. (8) is very small. This means that the probability of each individual person spreading an idea (or rumour or disease) need not be large for it still to spread through the whole community.

Another important class of network models is the class of generative models, models that posit a quantitative mechanism or mechanisms by which a network forms, usually as a way of explaining how the observed structure of the network arises. The best-known example of such a model is the 'cumulative advantage' or 'preferential attachment' model (Price 1976; Barabási and Albert 1999), which aims to explain the fat-tailed degree distributions observed in some networks. In its simplest form this model envisages a network that grows by the steady addition of vertices, one at a time. Many networks, such as the World Wide Web and citation networks, grow this way; it is a matter of current debate whether the model applies to social networks as well. Each vertex is added with a certain number m of edges emerging from it, whose other ends connect to pre-existing vertices with probability proportional to those vertices' current degree. That is, the higher the current degree of a vertex, the more likely that vertex is to acquire new edges when the graph grows. This kind of rich-get-richer phenomenon is plausible in many network contexts and is known to generate Pareto degree distributions. Using a rate-equation method (Price 1976; Simon 1955; Krapivsky et al. 2000), we find that in the limit of large network size the degree distribution obeys:

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}. \tag{9}$$

This distribution has a tail going as $p_k : k^{-3}$ in the large-k limit, which is strongly reminiscent of the degree distributions seen particularly in citation networks and also in the World Wide Web. Generative models of this type have been a source of considerable interest in recent years and have been much extended by a number of authors (Dorogovtsev and Mendes 2003; Albert and Barabási 2002) beyond the simple ideas described here.

Concepts such as those appearing in this article can be developed a great deal further and lead to a variety of useful, and in some cases surprising, results about the function of networked systems. More details can be found in the references.

See Also

- ▶ Artificial Neural Networks
- ▶ Business Networks
- ▶ Graph Theory
- ▶ Interacting Agents in Finance
- ▶ Network Formation
- ▶ Pareto Distribution
- ▶ Perron–Frobenius Theorem
- ▶ Power Laws
- ▶ Psychology of Social Networks
- ▶ Small-World Networks
- ▶ Social Networks in Labour Markets

Bibliography

Ahuja, R., T. Magnanti, and J. Orlin. 1993. *Network flows: Theory, algorithms, and applications*. Upper Saddle River: Prentice Hall.

Albert, R., and A.-L. Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74: 47–97.

Barabási, A.-L., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286: 509–512.

Barnes, J. 1954. Class and committees in a Norwegian island parish. *Human Relations* 7: 39–58.

Bollobás, B. 2001. *Random graphs*. 2nd ed. New York: Academic Press.

Davis, A., B. Gardner, and M. Gardner. 1941. *Deep south*. Chicago: University of Chicago Press.

Dorogovtsev, S., and J. Mendes. 2003. *Evolution of networks: From biological nets to the internet and WWW*. Oxford: Oxford University Press.

Erdős, P., and A. Rényi. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17–61.

Harary, F. 1995. *Graph theory*. Cambridge: Perseus.

Krapivsky, P., S. Redner, and F. Leyvraz. 2000. Connectivity of growing random networks. *Physical Review Letters* 85: 4629–4632.

Milgram, S. 1967. The small world problem. *Psychology Today* 2: 60–67.

Moreno, J. 1934. *Who shall survive?* Beacon: Beacon House.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45: 167–256.
Pool, I., and M. Kochen. 1978. Contacts and influence. *Social Networks* 1: 1–48.
Price, D. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27: 292–306.
Rapoport, A., and W. Horvath. 1961. A study of a large sociogram. *Behavioral Science* 6: 279–291.
Scott, J. 2000. *Social network analysis: A handbook*. 2nd ed. London: Sage.
Simon, H. 1955. On a class of skew distribution functions. *Biometrika* 42: 425–440.
Solomonoff, R., and A. Rapoport. 1951. Connectivity of random nets. *Bulletin of Mathematical Biophysics* 13: 107–117.
Wasserman, S., and K. Faust. 1994. *Social network analysis*. Cambridge: Cambridge University Press.
West, D. 1996. *Introduction to graph theory*. Upper Saddle River: Prentice Hall.

Matrix Multiplier

J. R. N. Stone

The matrix multiplier is a generalization of Kahn's scalar multiplier. The term was introduced by Goodwin (1949), though similar work was done independently by Chipman (1949, 1950a, 1951).

Consider a closed set of accounts, divided into endogenous and exogenous subsets and represented by the row-and-column pairs in a matrix with money incomings in the rows and money outgoings in the columns. The first n accounts are endogenous and the remaining m are exogenous. The $n \times n$ submatrix, W say, contains transactions between endogenous accounts; the $n \times m$ submatrix, X say, contains expenditures by the exogenous accounts which are injections into the endogenous subsystem; the $m \times n$ submatrix, Z say, contains expenditures by the endogenous accounts which are leakages into the exogenous subsystem.

Writing y for the column vector of endogenous account totals, i for the unit vector, I for the unit matrix, x for Xi, and $A = W\hat{y}^{-1}$ where $\hat{y}$ denotes a diagonal matrix formed from y, then

$$y = Wi + x = Ay + x = (I - A)^{-1}x \quad (1)$$

where $(I-A)^{-1}$ is a matrix multiplier transforming the injections into the endogenous accounts into the endogenous account totals. If these accounts are restricted to the production accounts of the economy, then (1) represents Leontief's open input–output model.

Since the entries in the matrix are expressed in money terms, the vector of product prices is $p = \hat{y}^{-1}y = i$; and if we denote the vector of exogenous outgoings (primary input costs) per unit of total outgoings from the endogenous accounts (total costs = total output) by $f = \hat{y}^{-1}z$, where $z = Z'i$, then

$$y = W'i + z \quad (2)$$

and

$$p = A'p + f = (I - A')^{-1}f \quad (3)$$

where $(I - A')^{-1}$, the transpose of $(I - A)^{-1}$, transforms the cost of primary inputs per unit of output into product prices. These multipliers exist only if $A'i < i$.

Since $A^{\theta} \to 0$ as $\theta \to \infty$, we can write

$$\begin{aligned}(I - A)^{-1} &= I + A + A^2 \ldots \\ &= I + A + A^2(I - A)^{-1} \quad (4)\end{aligned}$$

where I corresponds to the output required to meet the injections of demand; A corresponds to the direct inputs into this output; and $A^2(I-A)^{-1}$ corresponds to the indirect inputs required to produce the direct outputs and so on indefinitely.

The endogenous subsystem need not be restricted to production accounts. Pyatt, Roe and associates (1977) treat as endogenous: (i) accounts for different types of factor income, (ii) accounts for the income and outlay of different sectors, and (iii) accounts for different branches of production, leaving capital and foreign accounts as exogenous. As a result, A is partitioned into three diagonal submatrices, B say, and six off-diagonal ones,

C say. Then, putting $(I - B)^{-1}C = D$, we can write

$$\begin{aligned} y &= Ay + x = (I - D)^{-1}(I - B)^{-1}x \\ &= \left(I + D + D^2\right)\left(I - D^3\right)^{-1}(I - B)^{-1}x \\ &= M_3M_2M_1x. \end{aligned} \tag{5}$$

In (5), $(I - A)^{-1}$ is partitioned into three components: M_1 arises from repercussions of the initial injection within the subsystem it initially entered; M_2 arises from its repercussions when it has completed a tour through all three subsystems and returned to the one it initially entered; and M_3 arises from its repercussions when it has completed a tour outside its original subsystem without returning to it. Thus, in terms of additive components,

$$\begin{aligned} (I - A)^{-1} = I + (M_1 - I) + (M_2 - I)M_1 \\ + (M_3 - I)M_2M_1. \end{aligned} \tag{6}$$

The multipliers are based on average propensities, which should be replaced by marginal propensities when measuring the effects of changes in injections (Pyatt and Round 1979).

Equation (5) can be extended to several regions (Round 1984), though difficulties arise if we try to apply the analysis to more than three. For two regions, r and s say, let the diagonal matrices of transactions between them be $\hat{a}_{rs}$ and $\hat{a}_{sr}$. Then, putting $(I - A_{rr})^{-1}\hat{a}_{rs} = F_{rs}$, (5) is replaced by

$$\begin{aligned} \begin{bmatrix} zy_r \\ y_s \end{bmatrix} &= \begin{bmatrix} A_{rr} & \hat{a}_{rs} \\ \hat{a}_{sr} & A_{ss} \end{bmatrix}\begin{bmatrix} y_r \\ y_s \end{bmatrix} + \begin{bmatrix} x_r \\ x_s \end{bmatrix} \\ &= \begin{bmatrix} I & -F_{rs}F_{sr} \\ -F_{sr}F_{rs} & I \end{bmatrix}^{-1}\begin{bmatrix} (I - A_{rr})^{-1} & 0 \\ 0 & (I - A_{ss})^{-1} \end{bmatrix} \\ &\times \begin{bmatrix} x_r \\ x_s \end{bmatrix} = \begin{bmatrix} (I - F_{rs}F_{sr})^{-1} & 0 \\ 0 & (I - F_{sr}F_{rs})^{-1} \end{bmatrix} \\ &\times \begin{bmatrix} I & F_{rs} \\ F_{sr} & I \end{bmatrix} \times \begin{bmatrix} (I - A_{rr})^{-1} & 0 \\ 0 & (I - A_{ss})^{-1} \end{bmatrix}\begin{bmatrix} x_r \\ x_s \end{bmatrix} \\ &= N_3N_2N_1x \end{aligned} \tag{7}$$

where $x = \{x_r\ x_s\}$. The complete matrix multiplier for the two region-system can be expressed in terms of additive components by substituting Ns for Ms in (6).

Goodwin (1949) also formulates a dynamic matrix multiplier corresponding to (1). An economy is initially in equilibrium, with $y = y_0$ and $x = x_0$. It is assumed that there is a one-period lag in all spending and that in each period $\theta > 0, x_0$ is replaced by x_1. Writing $\Lambda^\theta x(\tau) = x(\tau + \theta)$,

$$\begin{aligned} \Lambda^\tau y = A^\tau y_0 + \sum_{\theta=0}^{\tau-1} A^\theta \Lambda^{\tau-\theta-1}x_1 \\ = A^\tau y + (I - A^\tau)(I - A)^{-1}x_1 \\ = A^\tau y_0 + (I - A^\tau)y_i \end{aligned} \tag{8}$$

that is, y moves from y_0 to y_1 as a changing weighted sum of the two. Thus the dynamic matrix multiplier has the static value as a limit.

Goodwin stated that successive values of the components of y were bound to oscillate and the question was debated in Chipman (1950b) and Goodwin (1950). In fact they oscillate only if the elements of $(x_1 - x_0)$ have different signs. Since this is likely to be the case, we can only work out the future course of y from the first row of (8), which would require estimates of the future course of injections.

See Also

- Linear Models
- Multiplier Analysis
- Perron–Frobenius Theorem

Bibliography

Chipman, J.S. 1949. The generalized bi-system multiplier. *Canadian Journal of Economics and Political Science* 15: 176–189.

Chipman, J.S. 1950a. The multi-sector multiplier. *Econometrica* 18: 355–374.

Chipman, J.S. 1950b. Professor Goodwin's matrix multiplier. *Economic Journal* 60: 753–763.

Chipman, J.S. 1951. *The theory of inter-sectoral money flows and income formation*. Baltimore: Johns Hopkins Press.

Goodwin, R.M. 1949. The multiplier as matrix. *Economic Journal* 59: 537–555.

Goodwin, R.M. 1950. Does the matrix multiplier oscillate? *Economic Journal* 60: 764–770.

Pyatt, G., and J.I. Round. 1979. Accounting and fixed price multipliers in a social accounting matrix framework. *Economic Journal* 89: 850–873.

Pyatt, G., Roe, A.R., and associates. 1977. *Social accounting for development planning with special reference to Sri Lanka*. Cambridge: Cambridge University Press.
Round, J.I. 1984. Decomposing multipliers for economic systems involving regional and world trade. *Economic Journal* 95: 383–399.

Maximum Likelihood

Jack R. Porter

Abstract

Maximum likelihood is a method of estimation developed for fully specified parametric likelihood settings. In smooth parametric models, maximum likelihood has a number of desirable properties, including consistency, asymptotic normality, and asymptotic efficiency. Maximum likelihood has been usefully extended to various semiparametric and nonparametric settings.

Keywords

Asymptotic normality; Bootstrap; Confidence region; Consistency; EM algorithm; Empirical likelihood; Fisher, R. A.; Generalized method of moments; Invariance; Law of large numbers; Likelihood principle; Local likelihood; Log likelihood ratio; Maximum likelihood; Nonparametric regression; Semiparametric estimation; Statistical inference; Sufficiency

JEL Classifications

C1

Given data from some member of a parametric family of distributions, maximum likelihood provides a general purpose method of estimation frequently accompanied by useful statistical properties.

In a series of papers, R.A. Fisher (1922, 1925, 1934) proposed and argued for a method of estimation he dubbed 'maximum likelihood'. The intuitive appeal and broad applicability continue to drive its use as a primary tool of statisticians. Suppose data $z = (z_1, \ldots, z_n)$ is drawn from a distribution with density $f_n(z; \theta_0)$, and further suppose that this distribution is a member of a family of parametric distributions with densities $\{f_n(z; \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ (for k finite and $\theta_0 \in \Theta$). The likelihood function is simply defined by the joint density as the function $l_n(\theta, z) = f_n(z; \theta)$ with argument θ and data z held fixed. The maximum likelihood estimator (MLE) is then defined as

$$\hat{\theta}_{ML} = \arg \min_{\theta \in \Theta} l_n(\theta, z).$$

One motivation for the MLE comes from the likelihood principle, which implies that statistical inference on θ given data z should be based solely on the likelihood function, $l_n(\theta, z)$ (Berger and Wolpert 1988). According to this principle the relative evidence on two different values of θ given by the data is fully summarized by their likelihood ratio. In this sense, the MLE is the value of θ most supported by the data. Of course, most econometric work involving maximum likelihood does not take a strict likelihood principle viewpoint, but is typically more concerned with sampling properties from a frequentist viewpoint. It is this perspective that will be our main emphasis in what follows.

It is often convenient to (equivalently) think of the MLE as maximizing the log likelihood ratio, $\mathcal{L}_n(\theta) = \ln \frac{f_n(z;\theta)}{f_n(z;\theta_0)}$. When the data is independent and identically distributed (i.i.d.) with marginal density f, the log likelihood ratio can be written $\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \ln \frac{f(z_i;\theta)}{f(z_i;\theta_0)}$. By the law of large numbers, the normalized log likelihood ratio $\left(\frac{1}{n}\mathcal{L}_n(\theta)\right)$ approaches $\mathcal{L}(\theta) = E\theta_0\left[\ln \frac{f(z_i;\theta)}{f(z_i;\theta_0)}\right]$ (where the expectation is taken with respect to the 'population' density $f(z_i; \theta_0)$) asymptotically. Though–$\mathcal{L}(\theta)$ does not satisfy the formal definition of metric, it is often taken as distance measure between the densities $f(z_i; \theta)$ and $f(z_i; \theta_0)$. Not surprisingly, this distance is minimized at $\theta = \theta_0$ (when the identification condition that $f(z; \theta) \neq f(z; \theta_0)$ for $\theta \neq \theta_0$ is satisfied). The log likelihood ratio $\mathcal{L}_n(\theta)$, can be interpreted as a sample approximation to

this discrepancy measure, which is minimized at the MLE. The likelihood ratio test statistic, for testing the null hypothesis that $\theta = \theta_0$, is also based on this value.

Fisher emphasized the usefulness of the maximized likelihood itself. The density $f_n\left(z; \hat{\theta}_{ML}\right)$ provides an approximation to the population density. If, for instance, there is interest in some feature of $f_n(z; \theta_0)$, then an approximation can often be obtained from the corresponding feature $f_n\left(z; \hat{\theta}_{ML}\right)$ (as in the parametric bootstrap). More generally, Efron (1982) notes that $f_n\left(z; \hat{\theta}_{ML}\right)$ acts as a data summary.

Properties

For most commonly used parametric distributional families, the MLE is consistent. Note, for instance, that when $\mathcal{L}(\theta)$ is maximized at θ_0, and the convergence of the log likelihood ratio $\frac{1}{n}\mathcal{L}(\theta)$ mentioned above is uniform on Θ, then $\hat{\theta}_{ML}$ will correspondingly converge to θ_0. More general sufficient conditions for consistency are also available; see Ibragimov and Has'minskii (1981).

Under appropriate regularity conditions (which essentially amount to smoothness of the parametric model), the MLE is asymptotically normal.

$$\sqrt{n}\left(\hat{\theta}_{ML} - \theta_0\right) \rightarrow N\left(0, J(\theta_0)^{-1}\right)$$

where $J(\theta)$ is the Fisher information matrix, with value $E_\theta[\nabla_\theta lnf(z;\ \theta)(\nabla_\theta lnf(z;\ \theta))']$ when the data is i.i.d. and $\nabla_\theta \ \ln f(z_i;\ \theta)$ is called the 'score function'. Define the Hessian as $H(\theta) = \ E_\theta[\nabla_{\theta\theta} \ \ln \ f(z_i;\ \ \theta)]$. By the information matrix equality, $J(\theta) = \ -H(\theta)$ which adds to the variety of estimators for the information matrix. Frequentist confidence intervals are then immediately available based on the asymptotic normality property and an estimator for $J(\theta_0)$.

Other approximations for the distribution of the MLE are also available for certain statistical models. Barndorff-Nielsen (1983) provides an accurate approximation to the conditional distribution of the MLE given a maximal ancillary statistic. (An ancillary statistic is a statistic whose distribution does not depend on θ, and if every ancillary statistic is a function of a given ancillary statistic then that statistic is called maximal.) When the MLE is sufficient, Barndorff-Nielsen's formula is exact. For non-regular models, asymptotic normality may no longer hold and a general limiting distribution result is then unavailable. Ibragimov and Has'minskii (1981), for instance, characterize the asymptotic behavior of the MLE for certain non-regular classes of models.

The Fisher information matrix is additionally useful as an efficiency bound. Accordingly, the MLE itself enjoys certain optimality properties. For regular models, the MLE is asymptotically efficient under classical criteria. Hirano and Porter (2005) show that a shifted version of the MLE is asymptotically efficient for an even broader class of statistical models (and allow for asymmetric loss). Higher-order efficiency of the MLE has been established in Pfanzagl and Wefelmeyer (1978).

Intuition for the asymptotic normality and efficiency of the MLE can be gained through a consideration of the behaviour of the log likelihood ratio in the i.i.d. case. If we re-parametrize the likelihood in terms of the 'local' parameter $h = \sqrt{n}(\theta - \theta_0)$, then with enough smoothness the log likelihood ratio can be expanded as follows

$$\sum_{i=1}^{n} \ln \frac{f(z_i; \theta_0 + h/\sqrt{n})}{f(z_i; \theta_0)} \approx \frac{h'}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \ln f(z_i; \theta_0) + \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} h' \nabla_{\theta\theta} \ln \ f(z_i; \theta_0) h.$$

Under regularity conditions, the log likelihood ratio converges in distribution (for each h) to $N\left(-\frac{1}{2}h'J(\theta_0)h, h'J(\theta_0)h\right)$. (This kind of 'Taylor' expansion actually holds under a mild condition of differentiability in quadratic mean which is weaker than the twice continuous differentiability of the likelihood that appears necessary.) Models with log likelihood ratios obeying this kind of convergence are called 'locally asymptotically normal'. Now, consider the statistical model consisting of a single observation on a random variable $X \sim N(h, J(\theta_0)^{-1})$. Notably, the log

likelihood ratio for this simple statistical model $\{N(h, J(\theta_0)^{-1}) : h \in \mathbb{R}^k\}$ has the same distribution as the asymptotic distribution for the log likelihood ratio of the general model above. Since the log likelihood ratio captures all the statistical information in a given statistical model, there is an equivalence between the asymptotic behaviour of the original model with densities $\prod_{i=1}^{n} f(z_i; \theta)$ and the much simpler model given by a single observation from a normal with unknown mean h (and known variance-covariance, $J(\theta_0)^{-1}$). This equivalence is formalized in the limits of experiments theory (Le Cam 1986). Intuitively, one might expect that the MLE for the local parameter $\hat{h}_{ML} = \sqrt{n}\left(\hat{\theta}_{ML} - \theta_0\right)$ in the original model will behave (asymptotically) like the MLE of the 'limit' normal model, which is simply given by X. The normality of X and the efficiency (minimax) of the mean in a normal model then corresponds to the asymptotic normality and asymptotic efficiency of the MLE in the original model.

Other important properties of the MLE are invariance and sufficiency. The MLE is necessarily a function of all sufficient statistics. The MLE is also invariant to parametrization of the family of distributions. So, if the distributions are reparametrized in terms of $\lambda = T(\theta)$, then $\hat{\lambda}_{ML} = T\left(\hat{\theta}_{ML}\right)$. Additionally, the MLE satisfies a group equivariance property (Eaton 1989). Suppose the family of distributions is invariant under the group of transformations $\mathcal{G}$ defined on both the sample and parameter spaces. if $g \in \mathcal{G}$, then $\hat{\theta}_{ML}(gz) = g\,\hat{\theta}_{ML}(z)$ where the MLE is written as a function of the observations.

Limitations

Since densities are not uniquely defined, the likelihood criterion on which the MLE is based is not uniquely defined. For a given likelihood, a solution to the maximization problem that defines the MLE need not necessarily exist (or multiple solutions are also possible).

The consistency, asymptotic normality and efficiency properties (discussed above) are all asymptotic, leaving the possibility that the small sample behaviour of the MLE may be quite poor in given applications. Even the asymptotic properties themselves are assured under regularity conditions. Neyman and Scott (1948) describe a famous example where maximum likelihood can be poorly behaved in small samples. The random variables x_{ij} are distributed $N(\mu_i, \sigma^2)$ for $i = 1,\ldots, n, j = 1,\ldots, J$, and all random variables are independent. Consider the case with fixed n and $J = 2$. Since $x_{i2} - x_{i1}$ is distributed $N(0, 2\sigma^2)$, $s_n^2 = \frac{1}{2}\ \frac{1}{n}\sum_{i=1}^{n}(x_{i2} - x_{i1})^2$ is a natural and reasonable estimator for σ^2. But $\hat{\sigma}^2{}_{ML} = \frac{1}{2}s_n^2$, which could be a quite poor estimator with significant bias. This poor small sample performance is particularly notable, since this model consists only of independent normally distributed random variables. Asymptotically, if n remains fixed and J grows, then the MLE has all the usual favourable large sample properties. If J is fixed and n grows, then the assumption of a finite dimensional parameter space is violated, and the MLE is not even consistent.

Stein's well-known shrinkage estimator shows that, even in a simple normal model with known variance-covariance and unknown mean, the MLE need not be (mean-squared error) optimal. It is also notable that, outside of regular models, asymptotic efficiency of the MLE can frequently fail. A simple example of such a non-regular model is data drawn from a uniform distribution on $[0, \theta]$. More general, parameter-dependent support models can be found in the auction literature, and the MLE is generally suboptimal by traditional asymptotic efficiency criteria (Hirano and Porter 2003). Le Cam (1990) lists a number of additional examples where the deficiencies of maximum likelihood are highlighted.

Extensions

Suppose the parameter is partitioned, $\theta' = (\theta'_1, \theta'_2)$ and we define $\theta_2^*(\theta_1) = \arg\max_{\theta 2} l_n(\theta_1, \theta_2)$. Then, the profiled likelihood, $l_n\left(\theta_1, \theta_2^*(\theta_1)\right)$ can be maximized to give the MLE for θ_1. Sometimes this is useful for computational purposes.

This formulation has also been useful for conceptual purposes, such as developing semiparametric efficiency bounds, where θ_2 contains nuisance parameters. Maximum likelihood theory also extends immediately to conditional likelihood formulations. Other methods have been developed to ease the computational burden of maximum likelihood in certain problems. The EM algorithm can be especially helpful in missing data cases (MacLachlan and Krishnan 1997). Simulated maximum likelihood is useful when the likelihood can be expressed as high-dimensional integral without a closed form solution (Hajivassiliou and Ruud 1994).

A natural concern with maximum likelihood is its reliance on correct specification of the family of distributions. Quasi-likelihood methods suggest parametric families that have robustness properties beyond the family specified. Exponential linear families often play a prominent role in this approach (Gourieroux et al. 1984). Typically, efficiency is sacrificed, but consistency and asymptotic normality still hold where the asymptotic variance of the limiting normal distribution is given by the 'sandwich' formula, $H(\theta_0)^{-1}J(\theta_0)^{-1}H(\theta_0)^{-1}$.

Extensions of maximum likelihood have also been usefully applied in semiparametric and nonparametric contexts. Ai (1997) considers semiparametric estimation in a model with unknown conditional density that is assumed only to satisfy an index restriction. The conditional density is estimated nonparametrically, and the corresponding score function is constructed to produce a semiparametric maximum likelihood estimate of a finite-dimensional parameter of the model. Tibshirani and Hastie (1987) introduced the notion of local likelihood estimation where regression functions are fit locally according to a maximum likelihood criterion. This idea has been extended to density estimation and other regression-type settings (Fan et al. 1998). Linton and Xiao (2007) develop an adaptive nonparametric regression approach that estimates the unknown density of the disturbance (or its score function) and then uses this estimate for local likelihood estimation of the unknown regression function. Empirical likelihood methods are another offshoot of nonparametric maximum likelihood. The basic insight that the empirical distribution function is the nonparametric MLE for a general cumulative distribution function has led to new approaches to confidence region formation, estimation in regression models, generalized method of moments inference, and bootstrapping (Owen 2001; Brown and Newey 2002).

See Also

- ▶ Classical Distribution Theories
- ▶ Econometrics
- ▶ Efficiency Bounds
- ▶ Empirical Likelihood
- ▶ Fisher, Ronald Aylmer (1890–1962)
- ▶ Non-parametric Structural Models
- ▶ Optimality and Efficiency
- ▶ Semiparametric Estimation
- ▶ Statistical Inference

Bibliography

Ai, C. 1997. A semiparametric maximum likelihood estimator. *Econometrica* 65: 933–963.

Barndorff-Nielsen, O.E. 1983. On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70: 343–365.

Berger, J., and R. Wolpert. 1988. *The likelihood principle*. Hayward: Institute of Mathematical Statistics.

Brown, B., and W. Newey. 2002. Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business and Economic Statistics* 20: 507–517.

Eaton, M. 1989. *Group invariance applications in statistics*. Hayward: Institute of Mathematical Statistics.

Efron, B. 1982. Maximum likelihood and decision theory. *Annals of Statistics* 10: 340–356.

Fan, J., M. Farmen, and I. Gijbels. 1998. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society (Series B)* 60: 591–608.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A* 222: 309–360.

Fisher, R.A. 1925. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22: 700–725.

Fisher, R.A. 1934. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London A* 144: 285–307.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo maximum likelihood methods: theory. *Econometrica* 52: 681–700.

Hajivassiliou, V., and P. Ruud. 1994. Classical estimation methods for LDV models using simulation. In *Handbook of Econometrics*, ed. D. McFadden and R. Engle, vol. 4. Amsterdam: North-Holland.
Hirano, K., and J. Porter. 2003. Asymptotic efficiency in parametric structural models with parameter dependent support. *Econometrica* 71: 1307–1338.
Hirano, K., and J. Porter 2005. *Efficiency in asymptotic shift models*, Working paper. University of Wisconsin.
Ibragimov, I.A., and R.Z. Has'minskii. 1981. *Statistical estimation: Asymptotic theory*. New York: Springer-Verlag.
Le Cam, L. 1986. *Asymptotic methods in statistical decision theory*. New York: Springer-Verlag.
Le Cam, L. 1990. Maximum likelihood: An introduction. *International Statistical Review* 58: 153–171.
Linton, O., and Z. Xiao. 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory* 23(3) (forthcoming)
MacLachlan, G., and T. Krishnan. 1997. *The EM algorithm and extensions*. New York: Wiley.
Neyman, J., and E.L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16: 1–32.
Owen, A.B. 2001. *Empirical likelihood*. New York: Chapman-Hall.
Pfanzagl, J., and W. Wefelmeyer. 1978. A Third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis* 8: 1–29.
Tibshirani, R., and T. Hastie. 1987. Local likelihood estimation. *Journal of the American Statistical Association* 82: 559–567.

Maximum Satisfaction

F. Y. Edgeworth

Maximum satisfaction is the object towards which the economic man strives; the margin, which constitutes economic equilibrium. A great part of economic theory may be regarded as a statement of the conditions of maximum satisfaction (cf. Marshall's *Principles*, mathematical appendix, note xiv). Thus the theory of market price – that the demand at that price should equal the supply at it (Mill, Bk. iii, ch. 2) – may be deduced as the condition of the price for which the satisfaction of the buyers and sellers should be a maximum.

It is understood that this maximum is subject, or – as the mathematicians say, *relative* – to certain limitations. Thus, in a market, it is assumed that property passes only by exchange. It is not denied that an equalization of property would – abstracting ulterior consequences – be productive of a greater sum total of utility than is produced by the play of the market under a regime of unequal property (Sidgwick's *Political Economy*, Bk. iii, ch. vii; Jevons's *Theory*, 2nd edn, p. 153; Marshall's *Principles*, Bk. v, ch. xiii).

It should be understood also that the *maximum* value of a function is not necessarily the greatest possible value, but only the greatest of all values in the neighbourhood – a peak, but not the summit. There may be more *maxima* than one; and one *maximum* may be greater than another (Marshall, *loc. cit*. note). Accordingly, while it is true that any disturbance by which trade is shifted from an equilibrium to a neighbouring position, causes a diminution in the sum total of utility, it is also true that a disturbance, by which trade is shifted to the neighbourhood of a new equilibrium, may cause an increase in the sum total of utility. The latter kind of change is apt to occur when, by a stimulus to increased production, the advantages of production on a large scale are secured. Now it is quite conceivable that such a stimulus should be given by governmental interposition. Thus, while it is right to hold with Auspitz and Lieben (*Theorie der Preises*, p. 425) and the classical economists, that a bounty causes a diminution in the sum total of utility, the organization of industry being supposed unchanged; it is also right to hold with Professor Marshall that bounty, by bringing about a re-organization of industry, may cause an increase in the sum total of utility.

Altogether, the doctrine that maximum satisfaction, or the greatest general good, is attained by exchange free from government intervention, is theoretically true in a much narrower sense than has been supposed by many publicists, and even by some theoretical economists. Its validity as a handy rule for practice is not denied.

[There is implied in the preceding argument a certain conception, which it is impossible here to express fully, concerning the modification of the law of supply – or Supply-curves – which is involved in a re-organization of industry,

consequent on an enlarged scale of production. The view expressed on this subject by H. Cunynghame in the *Economic Journal* for March 1892 may be compared with the view expressed by Edgeworth in the *Economic Journal*, vol. iv, p. 436, and vol. xv, p. 62. In his *Mathematical Psychics*, he has pointed out analogies between the principle of maximum utility in economics and the principle of maximum energy in physics.]

Bibliography

Auspitz, R., and R. Lieben. 1887. *Zur Theorie des Preises*. Leipzig: Duncker & Humblot.

Cunynghame, H. 1892. Some improvements in simple geometrical methods of treating exchange values, monopoly and rent. *Economic Journal* 2(1): 35–52.

Edgeworth, F.Y. 1881. *Mathematical Psychics*. London: Kegan Paul.

Edgeworth, F.Y. 1894. The theory of international values, Pt 2. *Economic Journal* 4(3): 424–443.

Edgeworth, F.Y. 1905. Review of *History of the English Corn Laws* by J.S. Nicholson. *Economic Journal* 15(1): 60–62.

Edgeworth, F.Y. 1925. *Papers relating to political economy*. 3 vols. London: Macmillan.

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

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

Marshall, A., and M. Paley. 1879. *The economics of industry*. London: Macmillan.

Mill, J.S. 1848. *Principles of political economy*. London: J.W. Parker.

Sidgwick, H. 1883. *Principles of political economy*. London: Macmillan.

Sidgwick, H. 1885. *Scope and method of economic science*. London: Macmillan.

Maximum Score Methods

Robert P. Sherman

Abstract

This article describes some aspects of maximum score estimation of parameters of multinomial and, especially, binomial choice models. In the context of binomial choice models, strengths and weaknesses of the estimation procedure are discussed, as well as its relation to classical quantile regression estimation and its nonstandard rate of convergence. The benefits of smoothing the score criterion function are also noted.

Keywords

Binary response models; Bootstrap; Central limit theorems; Heteroskedasticity; Linear median regression; Maximum likelihood; Maximum score methods; Multinomial choice models; Quantile regression; Random utility maximization; Semiparametric estimation

JEL Classification

C14

In a seminal paper, Manski (1975) introduces the maximum score estimator (MSE) of the structural parameters of a multinomial choice model and proves consistency without assuming knowledge of the distribution of the error terms in the model. As such, the MSE is the first instance of a semiparametric estimator of a limited dependent variable model in the econometrics literature.

Maximum score estimation of the parameters of a binary choice model has received the most attention in the literature. Manski (1975) covers this model, but Manski (1985) focuses on it. The key assumption that Manski (1985) makes is that the latent variable underlying the observed binary data satisfies a linear α-quantile regression specification. (He focuses on the linear median regression case, where $\alpha = 0.5$.) This is perhaps an under-appreciated fact about maximum score estimation in the binary choice setting. If the latent variable were observed, then classical quantile regression estimation (Koenker and Bassett 1978), using the latent data, would estimate, albeit more efficiently, the same regression parameters that would be estimated by maximum score estimation using the binary data. In short, the estimands would be the same for these two estimation procedures.

Assuming that the underlying latent variable satisfies a linear α-quantile regression specification is equivalent to assuming that the regression parameters in the linear model do not depend on the regressors and that the error term in the model has zero α-quantile conditional on the regressors. Under these assumptions, Manski (1985) proves strong consistency of the MSE. The zero conditional α-quantile assumption does not require the existence of any error moments and allows heteroskedastic errors of an unknown form. This flexibility is in contrast to many semiparametric estimators of comparable structural parameters for the binary choice model. As discussed in Powell (1994), many of these latter estimators require the existence of error moments and most require more restrictive assumptions governing the relation of errors to regressors.

The weak zero conditional α-quantile assumption comes at a price, however. Extrapolation power is limited: off the observed support of the regressors it is not possible to identify the conditional probability of the choice of interest, but only whether this probability is above or below $1 - \alpha$. See Manski (1995, pp. 149–50). There are also disadvantages associated with the estimation procedure. The maximum score criterion function is a sum of indicator functions of sets involving parameters. This lack of smoothness precludes using standard optimization routines to compute the MSE. Moreover, Kim and Pollard (1990) show that this type of discontinuity leads to a convergence rate of $n^{-1/3}$ rather than the $n^{-1/2}$ convergence rate attained by most semiparametric estimators of parameters in this model. In addition, Kim and Pollard (1990) show that the MSE has a nonstandard limiting distribution. The properties of this distribution are largely unknown, making asymptotic inference problematic. Also, Abrevaya and Huang (2005) prove that the bootstrapped MSE is an inconsistent estimator of the parameters of interest, precluding bootstrap inference.

To repair some of these shortcomings, Horowitz (1992) develops a smoothed MSE (SMSE) for the linear median regression case. This estimator retains the attractive flexibility properties of the MSE, but can be computed using standard optimization routines. In addition, the SMSE converges at a faster rate than the MSE and has a normal limit law allowing first order asymptotic inference. Horowitz (2002) proves that bootstrapped SMSE provides asymptotic refinements and in various simulations demonstrates the superiority of bootstrap tests over first-order asymptotic tests. Kordas (2006) generalizes Horowitz's (1992) SMSE to cover all α-quantiles.

In the next section, we present the multinomial choice model under random utility maximization as well as some intuition behind maximum score estimation in this context. We then discuss the relation between maximum score estimation in the binary response model and quantile regression. Next, we present Kim and Pollard's (1990) heuristic argument for the nonstandard rate of convergence of the MSE in the binary model. Finally, we discuss the method of Horowitz (1992) for smoothing the MSE.

The Random Utility Maximization Model of Choice and the MSE

Manski (1975) developed the MSE for the multinomial choice model in the context of random utility maximization. Suppose the ith individual in a sample of size n from a population of interest must make exactly one of J choices, where $J \geq 2$.

For $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, J\}$, let U_{ik} denote the utility to individual i of making choice j. Assume the structural form $U_{ij} = X'_{ij}\beta + \varepsilon_{ij}$ where X_{ij} is an observable $m \times 1$ vector of explanatory variables, β is a unknown $m \times 1$ parameter vector, and ε_{ij} is an unobservable random disturbance. (A more general set-up can be accommodated. For example, there can be a different parameter vector associated with each choice.)

The utilities associated with the choices an individual faces are latent, or unobservable. However, an individual's choice is observable. Suppose we adopt the maximum utility model of choice: if individual i makes choice j then $U_{ij} > U_{ik}$ for all $k \neq j$. For any event E, define the indicator function $\{E\} = 1$ if E occurs and 0 otherwise. Define

M

$$Y_{ij} = \{U_{ij} > U_{ik}, \quad \text{for all } k \neq j\}$$
$$= \left\{X'_{ij} + \varepsilon_{ij} > X'_{ij}\beta + \varepsilon_{ik}, \quad \text{for all } k \neq j\right\}. \tag{1}$$

If choice j has maximum utility, then $Y_{ij} = 1$. Otherwise, $Y_{ij} = 0$. Thus, for each individual i, we observe $X_{ij}, j = 1, 2, \ldots, J$,and Y_{ij}, $j = 1, 2, \ldots, J$.

The traditional approach to estimating β in the multinomial choice model under the assumption of random utility maximization is the method of maximum likelihood in which the errors are iid with a distribution known up to scale. The likelihood function to be maximized has the form

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij} \quad \log P\{Y_{ij} = 1|X_{i1}, X_{i2}, \ldots, X_{iJ}, b\}.$$

For example, when εij has the Type 1 extreme-value cdf $F(t) = \exp(-\exp(-t)), \quad t \in R$, McFadden (1974) shows that the likelihood probabilities have the multinomial logit specification $\exp\left(X'_{ij}b\right)\left[\sum_{k=1}^{J}\exp\left(X'_{ij}b\right)\right]^{-1}$. The corresponding likelihood function is analytic and globally concave. Despite the consequent computational advantages, this specification makes very strong assumptions about the distribution of the errors. The MSE is consistent under much weaker assumptions about the errors. Manski (1975) only assumes that the disturbances ε_{ij} are independent and identically distributed (iid) across choices and independent but not necessarily identically distributed across individuals.

Write b for a generic element of the parameter space. It follows trivially from (1) that the infeasible criterion function

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}\left\{X'_{ij}b + \varepsilon_{ij} > X'_{ik}b + \varepsilon_{ik}, k \neq j\right\}.$$

attains its maximum value of n at $b = \beta$. Since, for each i, the disturbances ε_{ij} are iid variates, this suggests estimating β with the maximizer of the so-called score function

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}\left\{X'_{ij}b > X'_{ij}b, k \neq j\right\}.$$

A score for a parameter b is the number of correct predictions made by predicting Y_{ij} to be 1 whenever $X'_{ij}b$ exceeds $X'_{ij}b$ for all $k \neq j$. A maximizer of the score function is an MSE of β. The maximizer need not be unique.

The MSE in the Binary Choice Model and Quantile Regression

Now consider the binary model where $J = 2$. Define $Y_i = Y_{i1}$ (implying $Y_{i2} = 1 - Y$) and $X_i = X_{i1} - i2$. Then the score function in (2) reduces to

$$\sum_{i=1}^{n}\left[Y_i\{X'_ib > 0\} + 1(1 - Y_i)\{X'_ib < 0\}\right]. \tag{3}$$

Substitute $1 - \{X'_ib > 0\}$ for $1 - \{X'_ib < 0\}$ in (3) and expand each summand to see that maximizing (3) is equivalent to maximizing

$$S_n(b) = n^{-1}\sum_{i=1}^{n}(2Y_i - 1)\{X'_ib > 0\}. \tag{4}$$

Note that $Y_i = \{Y_i^* > 0\}$ where $Y_i^* = X'_i\beta + \varepsilon_i$ with $\varepsilon_i = \varepsilon_{i1} - \varepsilon_{i2}$. For ease of exposition, write (Y^*, Y, X, ε) for (Y_1^*, Y_1, X_1, ε_1) and x for an arbitrary point in the support of X. Thus, $Y = \{Y* > 0\}$ where $Y* = X'\beta + \varepsilon$.

Before proceeding further, we must consider what interpretation to give to the parameter β in the last paragraph. The interpretation depends on our assumptions. For example, if we assume that β does not depend on x and that for every x, $E[Y * |x] = x'\beta$, then β is such that the conditional mean of Y^* given $X = x$ is equal to $x'\beta$. However, if we assume that MED $(Y * |x) = x'\beta$, then β is such that the conditional median of Y^* given $X = x$ is equal to $x'\beta$. In general, the β satisfying the conditional mean assumption will be different from the β satisfying the conditional median assumption. Similarly, if we assume that for $a \neq 0{:}5$, the conditional α-quantile of Y^* given x is equal to $x'\beta$, then

this β will, in general, be different from the β satisfying the conditional median assumption.

With this in mind, for $\alpha \in (0,1)$, write $Q_\alpha(Y* | x)$ for the α-quantile of Y^* given $X = x$. Fix an $\alpha \in (0,1)$ and assume the linear α-quantile regression specification. That is, assume that for each x in the support of X, there exists a unique parameter β_α, depending on α but not on x, such that $Q_\alpha(Y* | x) = x'\beta_\alpha$. This implies a zero conditional α-quantile restriction on ε: $Q_\alpha(\varepsilon | x) = 0$ for all x.

For $\alpha \in (0,1)$, define

$$S_n^\alpha(b) = n^{-1}\sum_{i=1}^{n}[(2Y - 1_i) - (1-2\alpha)]\{X_i'b > 0\}. \tag{5}$$

Clearly, $S_n^{0.5}(b) = S_n(b)$ in (4). Assume that the linear α-quantile regression specification holds for some $\alpha \in (0,1)$. To see that it makes sense, under this assumption, to estimate β_α with the maximizer of $S_n^\alpha(b)$, consider $S^\alpha(b) = ES_n^\alpha(b)$

We see that

$$\begin{aligned} S^\alpha(b) &= E^X[E[(2Y-1) - 1(1-2\alpha)\{X'b > 0\}|X] \\ &= E^X\left[\left[2P\{-\varepsilon < X'\beta^\alpha\} - 1\right) - (1-2\alpha)\right]\{X'b > 0\}\right]. \end{aligned}$$

The linear α-quantile regression specification implies a zero conditional α-quantile restriction on ε: for all x, $P\{\varepsilon \le 0|x\} \le \alpha$ and $P\{\varepsilon \ge 0|x\} \ge \alpha$. Thus, $x'\beta_\alpha > 0$ if and only if $P\{-\varepsilon \le x\beta_\alpha|x\} \ge P\{-\varepsilon \le 0|x\} \ge 1 - \alpha$. Deduce that for each possible value of X, the term in outer brackets in the last expression is maximized at $b = \beta_\alpha$. It follows that $S^\alpha(b)$ is maximized at $b = \beta_\alpha$. The analogy principle (Manski 1988) prescribes using a maximizer of $S_n^\alpha(b)$ to estimate β_α.

The Nonstandard Convergence Rate

The summands of the criterion function in (5) depend on b only through indicator functions of sets. As such, each summand has a 'sharp edge', to use the terminology of Kim and Pollard (1990). These authors provide a beautiful heuristic for why estimators that optimize empirical processes with sharp-edge summands converge at rate $n^{-1/3}$, rather than the usual $n^{-1/2}$ rate. They decompose the sample criterion function into a deterministic trend plus noise. Then, for each possible parameter value, they consider how the trend and the noise compete for dominance. Only a parameter value for which the trend does not overwhelm the standard deviation of the noise has a fighting chance of being an optimizer. Sharp edges produce standard errors with nonstandard sizes leading to the nonstandard $n^{-1/3}$ rate. We now examine how their argument works for the MSE for a very simple model.

Assume the median regression specification for the model $Y = \{\beta - X - \varepsilon > 0\}$. Thus, $\beta_{0.5} = (\beta, 1)$ where the slope coefficient is known to equal -1 and the intercept β is the unknown parameter of interest. Assume that ε has median zero and is independent of X, so that the conditional median zero restriction is trivially satisfied. Also, assume that the distributions of X and ε have everywhere positive Lebesgue densities.

Refer to (4). Define $S(b) = ES_n(b) = E(2Y - 1)\{Xb > 0\}$. In the intercept example, $S(b) = E(2\{\varepsilon < \beta - X\} - 1)\{X < b\}$. Simple calculations show that

$$S(b) = 2\int_{-\infty}^{b} F_\varepsilon(\beta - t)f_x(t)dt - F_x(b)$$

where $F_\varepsilon(\cdot)$ is the cdf of ε, $f_x(\cdot)$ is the pdf of X, and $F_x(\cdot)$ is the cdf of X. Write $F_\varepsilon(\cdot)$ for the pdf of ε. Again, simple calculations show that

$$\begin{aligned} S'(b) &= 2E_\varepsilon(\beta - b)f_x(b) - f_x(b)S''(b) \\ &= 2F_\varepsilon(\beta - b)f_x'(b) - f_x(b)2f_\varepsilon(\beta - b) \\ &\quad - f_x'(b). \end{aligned}$$

By the median restriction, we see that $S'(\beta) = 0$ and $S''(\beta) = -2f_x(\beta)f_\varepsilon(0) < 0$. Thus, $S(b)$ is locally maximized at $b = \beta$. In fact, the given assumptions imply that $S(b)$ is globally and uniquely maximized at $b = \beta$. The MSE maximizes $S_n(b) - S_n(\beta)$. For each b, decompose $S_n(b) - S_n(\beta)$ into a sum of a deterministic trend and a random perturbation:

$$S_n(b) - S_n(\beta) = S(b) - S(\beta) + [S_n(b) - S_n(\beta) - [S(b) - S(\beta)]].$$

A Taylor expansion about β shows that for b near β, the trend $S(b) - S(\beta)$ is approximately quadratic with maximum value zero at $b = \beta$:

$$S(b) - S(\beta) \approx S''(\beta)(b - \beta)^2.$$

By a central limit theorem, for large n, the random contribution $S_n(b) - S_n(\beta) - [S(b) - S(\beta)]$ is approximately normally distributed with mean zero and variance σ_b^2/n where

$$\sigma_b^2 = E[(2Y - 1)[\{X < b\} - \{X < \beta\}]]^2 - [E(2Y - 1)[\{X < b\} - \{X < \beta\}]]^2.$$

For b near β, the second term is much smaller than the first. It is the first term that accounts for the sharp-edge effect. It equals

$$F_x(\beta) + F_x(b) - 2[F_x(\beta)\{b > \beta\} + F_x(b)\{b < \beta\}].$$

A Taylor expansion of both $F_x(b)$ terms about β shows that this term is approximately equal to $|b - \beta| f_x(\beta)$ for b near β. Thus, near β, the criterion function $S_n(b) - S_n(\beta)$ is approximately equal to a quadratic maximized at β, namely, $-c_1 (b - \beta)^2$ for $c_1 > 0$, plus a zero-mean random variable with standard deviation equal to $c_2 n^{-1/2} |b - \beta|^{1/2}$ for $c_2 > 0$. Values of b for which $-c_1 (b - \beta)^2$ is much bigger in absolute value than $c_2 n^{-1/2} |b - \beta|^{-1/2}$ have little chance of maximizing $S_n(b) - S_n(\beta)$. Rather, the maximizer is likely to be among those b values for which, for some $c > 0$,

$$(b - \beta)^2 \le c n^{-1/2} |b - \beta|^{1/2}.$$

Rearranging, we see that the maximizer is likely to be among the b values for which

$$|b - \beta| \le c n^{-1/3}.$$

This is the essence of the heuristic presented by Kim and Pollard (1990) for $n^{-1/3}$ convergence rates. These authors also note that, when criterion functions are smooth, the variance of the random perturbation usually has order $|b - \beta|^2$ (instead of $|b - \beta|$) which, by the same heuristic, leads to the faster $n^{-1/2}$ convergence rate.

Smoothing the MSE

In order to remedy some of the shortcomings of the MSE, Horowitz (1992) develops a smoothed maximum score estimator (SMSE) under a linear median regression specification for the latent variable in the binary model. He replaces the indicator function in (4) with a smooth approximation. His SMSE maximizes a criterion function of the form

$$n^{-1} \sum_{i=1}^{n} (2Y_i - 1) K\left(X_i' b \backslash \sigma_n\right)$$

where K is essentially a smooth cdf and σ_n approaches zero as the sample size increases. Thus, $K(X_i' b/\sigma_n)$ approaches the indicator function $\{X_i' b > 0\}$ as $n \to \infty$. By smoothing out the sharp-edge of the indicator function in (4), Horowitz is able to use Taylor expansion arguments to show that the SMSE, under slightly stronger conditions than those required for consistency of the MSE, converges at rate n^{δ} for $2/5 \le \delta < 1/2$ and has a normal limit. The exact rate of convergence depends on certain smoothness assumptions and satisfies an optimality property (see Horowitz 1993). The normality result makes it possible to do standard asymptotic inference with the SMSE. Horowitz (2002) shows that the bootstrapped SMSE provides asymptotic refinements.

Kordas (2006) applies the smoothing technique of Horowitz (1992) to the criterion function in (5) and obtains asymptotic results similar to those of Horowitz (1992) for any $\alpha \in (0, 1)$.

See Also

▶ Quantile Regression

Bibliography

Abrevaya, J., and J. Huang. 2005. On the bootstrap of the maximum score estimator. *Econometrica* 73: 1175–1204.

Horowitz, J.L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica* 60: 505–531.

Horowitz, J.L. 1993. Optimal rates of convergence of parameter estimators in the binary response model with weak distributional assumptions. *Econometric Theory* 9: 1–18.

Horowitz, J.L. 2002. Bootstrap critical values for tests based on the smoothed maximum score estimator. *Econometrica* 111: 141–167.

Kim, J., and D. Pollard. 1990. Cube root asymptotics. *Annals of Statistics* 18: 191–219.

Koenker, R., and G. Bassett Jr. 1978. Regression quantiles. *Econometrica* 46: 33–50.

Kordas, G. 2006. Smoothed binary regression quantiles. *Journal of Applied Econometrics* 21: 387–407.

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

Manski, C.F. 1985. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27: 313–333.

Manski, C.F. 1988. *Analog estimation methods in econometrics*. New York: Chapman and Hall.

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

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, ed. P. Zarembka. New York: Academic.

Powell, J.L. 1994. Estimation of semiparametric models. In *Handbook of econometrics*, vol. 4, ed. R. Engle and D. McFadden. Amsterdam: North-Holland.

Mazzola, Ugo (1863–1899)

F. Caffé

Keywords

Free trade; Marginalist theory; Mazzola, U.; Pantaleoni, M.; Price determination; Public finance

JEL Classifications

B31

Italian economist, born in Naples on 16 September 1863; died in Courmayeur on 14 August 1899. When he was just 20 years old, he graduated from the University of Naples, where his economic course had been largely based on the ideas of Francesco Ferrara. He did postgraduate research in Berlin, where he came in touch with Adolf Wagner and carried out research on behalf of the Italian Ministry of Agriculture into the problems associated with providing insurance for the working classes. At the age of 24 Mazzola was appointed to the Chair of Public Finance at the University of Pavia, where in 1896 he, with other economists, bought up the *Giornale degli Economisti* and transformed it into a centre of liberal thought. It was in this journal that the most eminent economists of the time, Maffeo Pantaleoni, Antonio De Viti De Marco, and Vilfredo Pareto, had their work published.

Mazzola believed fervently in the concept of free trade and he fought the protectionist trends which were threatening the country as it moved towards industrialization. In economic doctrine he was attracted to marginalist theory, as advocated by Jevons and Menger, and to its more comprehensive version in general equilibrium theory. Using these tools of analysis he wrote *I dati scientifici della finanza pubblica*, published in 1890, and thus made 'a lasting contribution' (to quote Pantaleoni) to the foundation of the theory of public finance. Mazzola was an expert on German fiscal theories, but he disagreed with the rather ambiguous way in which they differentiated between individual and collective aims. Mazzola stressed that, in his view, individual objectives are conditioned by public aims (defence, security, and so on) which can only be achieved by means of political cooperation. He believed that the provision of public welfare was necessary for the attainment of collective aims. So he analysed the characteristics of public welfare and then examined the process of price determination by means of the principles of maximization of utility – characteristics of the marginalist theory. In this way the phenomenon of fiscal theory was brought within the sphere of general economic analysis.

Selected Works

1886. *L'assicurazione degli operai nella scienza e nella legislazione germanica*. Rome: Botta.
1890. *I dati scientifici della finanza pubblica*. Rome: Loescher.
1895. *L'imposta progressiva in economia pura e sociale*. Pavia.
1967. The formation of the prices of public goods. In *Classics in the theory of public finance*, ed. R.A. Musgrave and A.T. Peacock. London: Macmillan.

Bibliography

Pantaleoni, M. 1899. Ugo Mazzola. *Giornale degli Economisti*. 2nd Series, 19: 189–198.

McCulloch, John Ramsay (1789–1864)

D. P. O'Brien

Keywords

Absolute advantage; Cobweb; Combination laws; Commodity money; Convertibility; Corn laws; Cost of production; Inflation; Invariable measure; Labour supply; Labour theory of value; McCulloch, J. R.; Poor law; Population theory; Public finance; Relative value; Rent; Transfer problem; Trade unions; Wages fund

JEL Classifications

B31

MuCulloch was born in Galloway, Scotland on 1 March 1789. After attending Edinburgh University he secured employment as a lawyer's clerk. In 1816 he began his contributions to economics with two essays on the national debt. He was editor of *The Scotsman*, 1817–21, and a contributor to that paper until 1827. In 1818 he began writing for the *Edinburgh Review* and continued doing so until 1837, contributing nearly 80 articles. He also contributed to *Encyclopaedia Britannica* and was a prolific author, his works including editions of the *Wealth of Nations*, a *Commercial Dictionary, a Geographical Dictionary, Principles of Political Economy,* a *Statistical Account of the British Empire* and a *Treatise on Taxation*. He was a noted bibliophile, and after his death his library was purchased by his friend Lord Overstone and ultimately presented to Reading University.

It was not only as an author that McCulloch was influential; he was also called as an expert witness before the Select Committee on machinery of 1824 and that on Ireland of 1825. He was also one of the first public teachers of economics. He began lecturing in Edinburgh in 1820, and although attempts to establish a chair at Edinburgh University on his behalf in 1825 failed, he had been selected in 1824 to give the Ricardo Memorial lectures in London. In 1828, largely through the agency of James Mill, he was appointed the first professor of political economy at London University. He remained professor until 1837, though largely supporting himself by his pen; but in 1838 he was appointed Comptroller of the Stationery Office, a post he occupied until his death in 1864. This did not prevent him from continuing his literary activities and he remained active as an author, producing new editions of his earlier works and some completely new ones, notably the *Treatise on Taxation* (1845, 1852, 1863) and the *Geographical Dictionary* (1841–2, 1845–6, 1849, 1852, 1854).

As an economist, McCulloch was at one stage very much under the influence of Ricardo, but the influence was transient. He put forward a simple labour theory of value under the influence of Ricardo, while aware that this was a simplified version (assuming away the problem of capital), for popular exposition. But though the emphasis is on labour quantity, and undoubtedly derived from Ricardo, it was a theory of relative (exchangeable) value, and McCulloch's analysis never borrowed from Ricardo the fundamental (to Ricardo)

concept of the invariable measure. Indeed, McCulloch rejected the concept emphatically.

The vicissitudes of McCulloch's exposition over the years included, without doubt, a number of erroneous positions and a strange solution to the origin of profits as being in stored labour continuing to work. He also attempted to make labour cost a real (disutility) cost. But in the second edition (1830) of his *Principles* he advanced an almost complete cost of production theory, where cost was, as in an earlier *Scotsman* article, marginal cost. Labour cost, amortization and profit were now recognized as costs as they had been with Smith; the influences of Ricardo was now largely apparent in treating rent not as a cost but as a price-determined surplus, thus ignoring transfer earnings. Finally, in the 1838 edition of his notes to the *Wealth of Nations*, McCulloch clarified the nature of the profit reward with both waiting (as stressed by Ricardo) and productivity (as stressed by McCulloch) combining to produce a positive return.

McCulloch's treatment of money and banking derived from Smith, Hume and Thornton, though he long followed Ricardo in advancing the idea that the value of commodity money was determined by its cost of production, an idea he finally abandoned in his contributions to the eighth edition of *Encyclopaedia Britannica*. He accepted Hume's theory of the distribution of the precious metals, drawing also on Smith and especially Thornton, though he did accept, at least in relation to external losses of metal, the Ricardian definition of excess. In considering the internal level of activity in relation to money, he accepted Say's identify as an equilibrium proposition; but he also recognized the possibility that excess demand for money in disequilibrium might cause economic dislocation, thus magnifying fluctuations originating in the real side of the economy, such as over-investment. He also recognized, and approved, in contrast to Ricardo, the effects of mild inflation in producing forced saving and economic growth. On the issue of banking control he at first accepted Ricardo's view that convertibility was its own safeguard, but he came to recognize the problems of over-issue of notes much earlier than many writers – he was firmly opposed to laissez-faire in banking – and was one of the earliest writers to put forward the principle that a note issue should fluctuate in amount in response to the balance of payments exactly as an identically circumstanced metallic money would do, though he saw this as providing only a partial solution to monetary control.

McCulloch's analysis of international trade followed Smith rather than Ricardo, in basing trade on absolute advantage assuming international factor mobility. McCulloch may have done this because he saw that the possibility of trade advantage, as explained in comparative cost theory, was incomplete until this was translated into relative costs and prices. At all events he considered Ricardo's treatment of international trade to be faulty. In his view there was a complete parallel between international and inter-regional trade. McCulloch's treatment went well beyond that of Smith in some respects; and in particular, his analysis of the transfer problem, based on the work of Parnell, was an important precursor of modern developments. He discussed not only the effects of a transfer in the form of specie, or of commodities, but also a demand transfer of the kind made famous by Ohlin in his controversy with Keynes after the First World War. On matters of trade policy McCulloch has had the image of a crude free trader: but in fact, though he recognized the harm that protection could do, freedom of trade could, in his view, involve the imposition of substantial import duties – even as high as 25 per cent – as long as these were balanced by home excise duties so as to avoid distortion of choice.

McCulloch's treatment of public finance used the Smithian framework: the analysis inevitably acquired a number of Ricardian accretions, but many of these were ultimately discarded. Moreover, McCulloch drew on a wide range of earlier writers on taxation and was particularly indebted to Hume and to Robert Hamilton. He presented a broad synthetic treatment which did much to give tax theory practicality after the Ricardian detour into the corn model. His main focus of attention was the use of fiscal policy in such a way as to

ensure the maintenance of growth. Heavy taxation could interfere with growth; but if taxation was sensibly used it could be a stimulus to growth, increasing the supply of both effort and savings. A widely based regime of moderate indirect taxes, extending even to postage, was what McCulloch favoured, on ability-to-pay grounds, despite the distortions of the price system and the regressive elements. He opposed Gladstone's taxation policy, not only reliance upon income tax which McCulloch believed to interfere with growth, falling to stimulate effort and, where not proportional, subverting economic motivation.

The basic process of economic development was seen largely in terms of Smith's apparatus involving the accumulation of capital (but including human capital) and division of labour. It also involved the institutional requirements of security of property, internal freedom of trade, and a substantial role for government including education, control of public utilities, and employer liability for accidents. On to this Smithian basis McCulloch grafted specific Ricardian features, notably the Ricardian explanation for the declining rate of profit. However, he finally rejected the idea of inevitably diminishing returns in agriculture, as well as the inverse movement of wages and profits, and the Ricardian stagnation thesis. Writing later than Smith, McCulloch's treatment of development shows a much heavier emphasis on technology. Indeed, this was the basis of a fundamental disagreement with Ricardo over the role of machinery, and led him also to reject the primacy which Smith had afforded to agriculture – he attached key importance to the manufacturing sector though he was worried about the distributional consequences.

On agriculture itself however he wrote a good deal. He believed in large-scale capitalist farming (and supported primogeniture though not entails), and came, after a long period of believing in the (at least ultimate) inevitability of diminishing returns, to the view that under such a system improvements might continuously offset the diminution of returns. Thus his attack on the Corn Laws did not emphasize the Ricardian concept of stagnation but other matters, notably the idea that they encouraged price fluctuation. In this context he employed not only arguments about elasticity of demand and supply but also an agricultural cobweb. On this basis he argued, in contrast to Ricardo, that all classes, including the landlords, lost by the Corn Laws (though he followed Ricardo's argument for a measure of agricultural protection) and he also came to reject Ricardo's argument that improvements were against the interest of the landlords.

Classical economics, unlike modern economics, contained, as an integral part, a theory of population, which served in turn as the basis for various theorems about wages and welfare. McCulloch's population theory initially followed the first two editions of Malthus's celebrated *Essay*. But, probably under the influence of Nassau Senior, he become opposed to the Malthusian argument, believing prudential restraints to predominate. Indeed, amongst mainstream classical economists he was probably the most extreme anti-Malthusian, a development which harmonized with his move away from Ricardo's influence. Having sided with Malthus and Ricardo in opposing the Poor Law, he changed his mind quite openly after 1826 and became a supporter of the old Poor Law, believing that it did not undermine prudential restraint and that it preserved social stability. He opposed the 1834 Poor Law as harsh and over-centralized. As a measure to raise wages generally he supported emigration and colonization, though he did not favour retention of control of colonies, and objected, in particular, to Wakefield's schemes for 'scientific' colonization.

All this raises directly McCulloch's concept of the operation of the labour market. McCulloch is particularly associated with the idea of the wage fund, because of his *Essay on Wages* (1826); and with a given wage fund, reducing supply will raise wages, as implied in McCulloch's treatment of emigration. His analysis of demand for labour thus equated capital with demand for labour, though in his more careful treatments he distinguished total and wage capital (and total population and labour supply). But this only provided half the analysis: on the supply side McCulloch employed four different labour supply functions,

including a rising supply schedule as normal; two negatively inclined short-run schedules (the first when wages rose after excessive hours had been required to survive at the previous level of wages, the second where women and children entered the labour force as wages fell, to maintain family income). Fourthly, there was a secular population function. McCulloch favoured high wages, and his writings in defence of trades unions were important in the successful struggle to secure repeal of the Combination Laws.

As an economist, McCulloch was a fairly representative classical writer in that his work involved a synthesis of elements deriving from Smith and Ricardo. Yet McCulloch's case is particularly interesting because his own evolution over they very long period (48 years) of his writings, mirrors the development of classical economics itself. Starting from a basis of Smith and Malthus, he fell under the influence of Ricardo's magnetic personality and remarkable powers of abstraction; but he gradually passed through this phase, the Smithian elements in his work resuming their predominance and leading in turn to an emphasis on empirical work and a methodological position foreign to the tenor of Ricardo's work.

Selected Works

1825. *The principles of political economy: With some inquiries respecting their application, and a sketch of the rise and progress of the science*, 3rd ed. Edinburgh: W. Tait. 1843.

1832. *A dictionary practical, theoretical, and historical, of commerce and commercial navigation*, 10th ed. London: Longmans. 1859.

1845. *A treatise on the principles and practical influence of taxation and the funding system*, ed. D.P. O'Brien. Edinburgh: Scottish Academic Press for The Scottish Economics Society. 1975.

Bibliography

O'Brien, D.P. 1970. *J.R. McCulloch: A study in classical economics*. London: Allen & Unwin.

McFadden, Daniel (Born 1937)

John Rust

Abstract

This article reviews the contributions of Daniel L. McFadden, 2000 co-recipient of the Nobel Prize in Economics. The article focuses on his seminal contributions to the theoretical and econometric literatures on discrete choice.

Keywords

Curse of dimensionality; Discrete choice; Duality; Generalized extreme value; Independence of irrelevant alternatives; Infinite horizons; Logit models of individual choice; Mathematical psychology; Maximum likelihood; McFadden, D. L.; Method of simulated moments; Monte Carlo methods; Multinomial logit model; Multinomial probit model; Neural networks; Overlapping generations models; Probabilistic choice theory; Random utility models; Revealed preference; Simulation-based estimation

JEL Classifications

B31

Introduction

Daniel L. McFadden, the E. Morris Cox Professor of Economics at the University of California at Berkeley, was the 2000 co-recipient of the Nobel Prize in Economics, awarded 'for his development of theory and methods of analyzing discrete choice'. (The prize was split with James J. Heckman, awarded 'for his development of theory and methods for analyzing selective samples'). McFadden was born in North Carolina, USA, in 1937 and received a BS in physics from the University of Minnesota (with highest honors)

in 1956, and a Ph.D. in economics from Minnesota in 1962. His academic career began as a postdoctoral fellow at the University of Pittsburgh. In 1963 he was appointed as assistant professor of economics at the University of California at Berkeley, and tenured in 1966. He has also held tenured appointments at Yale University (as Irving Fisher Research Professor in 1977), and at the Massachusetts Institute of Technology (from 1978 to 1991). In 1990 he was awarded the E. Morris Cox Chair at the University of California at Berkeley, where he has also served as Department Chair and as Director of the Econometrics Laboratory.

Research Contributions

McFadden is best known for his fundamental contributions to the theory and econometric methods for analysing *discrete choice.* Building on a highly abstract, axiomatic literature on *probabilistic choice theory* due to Thurstone (1927), Block and Marschak (1960), and Luce (1959), McFadden developed the econometric methodology for estimating the utility functions underlying probabilistic choice theory. McFadden's primary contribution was to provide the econometric tools that permitted widespread *practical empirical application* of discrete choice models, in economics and other disciplines. According to his autobiography (McFadden 2001),

> In 1964, I was working with a graduate student, Phoebe Cottingham, who had data on freeway routing decisions of the California Department of Transportation, and was looking for a way to analyze these data to study institutional decision-making behavior. I worked out for her an econometric model based on an axiomatic theory of choice behavior developed by the psychologist Duncan Luce. Drawing upon the work of Thurstone and Marshak, I was able to show how this model linked to the economic theory of choice behavior. These developments, now called the multinomial logit model and the random utility model for choice behavior, have turned out to be widely useful in economics and other social sciences. They are used, for example, to study travel modes, choice of occupation, brand of automobile purchase, and decisions on marriage and number of children.

Thousands of papers applying his technique have been published since his path-breaking papers, 'Conditional Logit Analysis of Qualitative Choice Behavior' (1973) and 'The Revealed Preferences of a Government Bureaucracy: Empirical Evidence' (1976). In December 2005, a search of the term 'discrete choice' using the Google search engine yielded 10,200,000 entries, and a search on the Google Scholar search engine (which limits search to academic articles) returned 759,000 items.

Besides the discrete choice literature itself, McFadden's work has spawned a number of related literatures in econometrics, theory, and industrial organization that are among the most active and productive parts of the economic literature in the present day. This includes work in game theory and industrial organization (for example, the work on discrete choice and product differentiation of Anderson et al. (1992), estimation of discrete games of incomplete information (Bajari et al. 2005), and discrete choice modelling in the empirical industrial organization literature (Berry et al. 1995, and Goldberg 1995), the econometric literature on semiparametric estimation of discrete choice models (Manski 1985; McFadden and Train 2000), the literature on discrete/continuous choice models and its connection to durable goods and energy demand modelling (Dagsvik 1994; Dubin and McFadden 1984; Hannemann 1984), the econometric literature on choice based and stratified sampling (Cosslett 1981; Manski and McFadden 1981), the econometric literature on 'simulation estimation' (McFadden 1994; Hajivassiliou and Ruud 1994; Pakes and Pollard 1989), and the work on structural estimation of dynamic discrete choice models and extensions thereof (Dagsvik 1983; Eckstein and Wolpin 1989; Heckman 1981; Rust 1994).

McFadden has also made significant contributions to other fields, particularly to economic theory and production economics. Due to space limitations, I can only briefly mention several of his best known contributions here. McFadden's earliest published work was in pure theory, including seminal work on duality theory of production functions that was subsequently published in his book on Production Economics edited with

Melvyn Fuss in 1978. McFadden made important contributions to growth theory including his 1967 Review of Economic Studies paper that showed how the overtaking criterion could be used to evaluate infinite horizon development programmes, resolving an outstanding paradox raised by Diamond and Koopmans. In a series of papers with Mitra and Majumdar (1976, 1980), McFadden extended the classical competitive equilibrium welfare theorems established by Debreu and others in finite economies, (that is, competitive equilibria are Pareto efficient, and any Pareto efficient allocation can be sustained as a competitive equilibrium after a suitable reallocation of resources), to infinite horizon economies. This work was not a simple technical extension or previous work by Debreu: it resolved serious conceptual problems created by the fact that in an infinite horizon economy (which includes standard overlapping generations models) the commodity space is infinite-dimensional and the number of consumers is infinite. These papers provided sufficient conditions for the existence of these fundamental welfare theorems, resolving paradoxes raised by Paul Samuelson, who showed special cases of infinite horizon overlapping generation economies where competitive equilibria can be strikingly inefficient. Another landmark paper is McFadden's (1974) paper on excess demand functions with Mantel, Mas-Colell and Richter. This paper provided one of the most general proofs of a classic conjecture by Hugo Sonnenschein that the necessary and sufficient properties of any system of aggregate excess demand functions are that it satisfy the following three properties: (1) homogeneity, (2) continuity, and (3) Walras's Law. McFadden has made numerous other contributions to economic theory that I do not have space to cover here.

Instead, I now return to a more in depth review of McFadden's contributions to the discrete choice literature, the primary contributions that were cited in his Nobel Prize award.

Contributions to Discrete Choice

McFadden's contributions built on prior work in the literature on *mathematical psychology* (see logit models of individual choice for further details). McFadden's contribution to this literature was to recognize how to operationalize the random utility interpretation in an empirically tractable way. In particular, he provided the first a random utility interpretation of the multinomial logit (MNL) model. His other fundamental contribution was to solve an analogue of the *revealed preference problem:* that is, using data on the actual choices and states of a sample of agents $\{(d_i, x_i)\}_{i=1}^{N}$, he showed how it was possible to 'reconstruct' their underlying random utility functions via the method of maximum likelihood, where the likelihood is a product of individuals' *conditional choice probabilities*. Given the simplicity of the MNL choice probabilities, this worked helped to spawn a huge empirical literature that applied discrete choice models to a wide variety of phenomena. Further, McFadden introduced a new class of multivariate distributions, the *generalized extreme value family* (GEV), and derived tractable formulas for the implied choice probabilities including the *nested multinomial logit model,* and showed that these models relax some of the empirically implausible restrictions implied by the multinomial logit model, particularly the *independence from irrelevant alternatives* (IIA) property.

Multivariate Extreme Value Distributions and the Multinomial Logit Model

McFadden assumed that an individual's utility function has the following *additive separable* representation

$$U(x, z, d, \theta) = u(x, d, \theta) + v(z, d). \qquad (1)$$

Define $\varepsilon(d) \equiv v(z, d)$. It follows that an assumption on the distribution of the random vector z implies a distribution for the random vector $\varepsilon \equiv \{\varepsilon(d) | d \in D(x)\}$. McFadden's approach was to make assumptions directly about the distribution of ε, rather than making assumptions about the distribution of z and deriving the implied distribution of ε. Standard assumptions for the distribution of ε that have been considered include the *multivariate normal* which yields the

multivariate probit variant of the discrete choice model. Unfortunately, in problems where there are more than only two alternatives (the case that Thurstone studied), the multinomial probit model becomes intractable in higher dimensional problems. The reason is that, in order to derive the conditional choice probabilities, one must do numerical integrations that have a dimension equal to $|D(x)|$, the number of elements in the choice set. In general this multivariate integration is computationally infeasible when $|D(x)|$ is larger than 5 or 6, using standard quadrature methods on modern computers.

McFadden introduced an alternative assumption for the distribution of ε, namely the *multivariate extreme value distribution* given by

$$F(z|x) = \Pr\{\varepsilon_d \le z_d | d \in D(x)\} = \prod_{d \in D(x)} \exp\{-\exp\{-(z_d - \mu_d)/\sigma)\}\}, \quad (2)$$

and showed that (when the location parameters μ_d are normalized to) the corresponding random utility model produces choice probabilities given by the multinomial logit formula

$$P(d|x,\theta) = \frac{\exp\{u(x,\ d,\ \theta)/\sigma\}}{\sum_{d' \in D(x)} \exp\{u(x,\ d',\ \theta)/\sigma\}}.$$

This is McFadden's key result, that is, the *MNL choice probability is implied by a random utility model when the random utilities have extreme value distributions.* It leads to the insight that the *independence from irrelevant alternatives* (IIA) property of the MNL model is a consequence of the statistical independence in the random utilities. In particular, even if the *observed attributes* of two alternatives d and d' are identical (which implies $u(x, d, \theta) = u(x, d', \theta)$), the statistical independence of unobservable components $\varepsilon(d)$ and $\varepsilon(d')$ implies alternatives d and d' are not perfect substitutes even when their observed characteristics are identical. In many cases this is not problematic: individuals may have different idiosyncratic perceptions and preferences for two different items that have the same observed attributes. However in the case of the 'red bus/blue bus' example or the concert ticket example discussed by Debreu (1960), there are cases where it is plausible to believe that the observed attributes provide a sufficiently good description of an agent's perception of the desirability of two alternatives. In such cases, the hypothesis that choices are also affected by additive, *independent* unobservables $\varepsilon(d)$ provides a poor representation of an agent's decisions. What is required in such cases is a random utility model that has the property that the degree of correlation in the unobserved components of utility $\varepsilon(d)$ and $\varepsilon(d')$ for two alternatives $d, d' \in D(x)$ is a function of the degree of closeness in the observed attributes. This type of dependence can be captured by a *random coefficient probit model.* This is a random utility model of the form $U(x, z, d, \theta) = x_d(\theta + z)$ where x_d is a $k \times 1$ vector of observed attributes of alternative d, and θ is a $k \times 1$ vector of utility weights representing the mean weights individuals assign to the various attributes in x_d in the population and $z \sim N(0,\Omega)$ is a $k \times 1$ normally distributed random vector representing agent specific deviations in their weighting of the attributes relative the population average values, θ. Under the random coefficients probit specification of the random utility model, when $x_d = x_{d'}$, alternatives d and d' are in fact perfect substitutes for each other and this model is able to provide the intuitively plausible prediction of the effect of introducing an irrelevant alternative – the red bus – in the red bus/blue bus problem (see, for example, Hausman and Wise 1978).

Generalized Extreme Value Distributions and Nested Logit Models

McFadden (1981) introduced the generalized extreme value (GEV) family of distributions. This family relaxes the independence assumption of the extreme value specification while still yielding tractable expressions for choice probabilities. The GEV distribution is given by

$$F(z|x) = \Pr\{\varepsilon_d \le z_d | d \in D(x)\} = \exp\{-H(\exp\{-z_1\}, \ldots, \exp\{-z_{|D(x)|}\}, x, D(x))\},$$

for any function $H(z, x, D(x))$ satisfying certain consistency properties. McFadden showed that choice probabilities for the GEV distribution are given by

$$P(d|x,\theta) = \frac{\exp\{u(x,d,\theta)\}H_d(\exp\{u(x,\ 1,\ \theta)\},\ldots,\exp\{u(x,|D(x)|,\theta)\},x,D(x))}{H(\exp\{u(x,1,\theta)\},\ldots,\exp\{u(x\ |D(x)|,\theta)\},x,D(x)),}$$

where $H_d(z, x, D(x)) = \partial/\partial_{Zd}\ H(z, x, D(x))$. A prominent subclass of GEV distributions is given by H functions of the form

$$H(z,y,D(x)) = \sum_{i=1}^{n}\left[\sum_{d\in D_i(x)} z_d^{1/\sigma_i}\right]^{\sigma_i},$$

where $\{D_1(x), \ldots, D_n(x)\}$ is a partition of the full choice set $D(x)$. This subclass of GEV distributions yields the *nested multinomial logit* (NMNL) choice probabilities (see logit models of individual choice for further details).

The NMNL model has been applied in numerous empirical studies especially to study demand where there are an extremely large number of alternatives, such as modelling consumer choice of automobiles (for example, Berkovec 1985; Goldberg 1995). In many of these consumer choice problems there is a natural partitioning of the choice set in terms of *product classes* (for example, luxury, compact, intermediate, sport-utility, and so on, classes in the case of autos). The nesting avoids the problems with the IIA property and results in more reasonable implied estimates of demand elasticities than those obtained using the MNL model. In fact, Dagsvik (1995) has shown that the class of random utility models with GEV distributed utilities is 'dense' in the class of all random utility models, in the sense that choice probabilities implied from any random utility model can be approximated arbitrarily closely by a random utility model in the GEV class. However a limitation of nested logit models is that they imply a highly structured pattern of correlation in the unobservables induced by the econometrician's specification of how the overall choice set $D(x)$ is to be partitioned, and the number of levels in the nested logit 'tree'. Even though the NMNL model can be nested to arbitrarily many levels to achieve additional flexibility, it is desirable to have a method where patterns of correlation in unobservables can be estimated from the data rather than being imposed by the analyst. Further, even though McFadden and Train (2000) recognize Dagsvik's (1995) finding as a 'powerful theoretical result', they conclude that 'its practical econometric application is limited by the difficulty of specifying, estimating and testing the consistency of relatively abstract generalized Extreme Value RUM' (McFadden and Train 2000, p. 452).

Method of Simulated Moments and Simulation Based Inference for Discrete Choice

As noted above, the random coefficients probit model has many attractive features: it allows a flexibly specified covariance matrix representing correlation between unobservable components of utilities that avoid many of the undesirable features implied by the IIA property of the MNL model, in a somewhat more direct and intuitive fashion than is possible via the GEV family. However as noted above, the multinomial probit model is intractable for applications with more than four or five alternatives due to the 'curse of dimensionality' of the numerical integrations required, at least using deterministic numerical integration methods such as Gaussian quadrature. One of McFadden's most important contributions was his (1989) *Econometrica* paper that introduced the *method of simulated moments* (MSM). This was a major breakthrough that introduced a new econometric method that made it feasible to estimate the parameters of multinomial probit models with arbitrarily large numbers of alternatives.

The basic idea underlying McFadden's contribution is to use *Monte Carlo integration* to approximate the probit choice probabilities. While this

idea had been previously proposed by others, it was never developed into a practical, widespread estimation method because 'it requires an impractical number of Monte Carlo draws to estimate small choice probabilities and their derivatives with acceptable precision' (McFadden 1989, p. 997). However McFadden's insight was that it is not necessary to have extremely accurate (and thus very computationally time-intensive) Monte Carlo estimates of choice probabilities in order to obtain an estimator for the parameters of a multinomial probit model that is consistent and asymptotically normal and performs well in finite samples. McFadden's insight is that the *noise from Monte Carlo simulations can be treated in the same way as random sampling error and will thus 'average out' in large samples.* In particular, his MSM estimator has good asymptotic properties even when *only a single Monte Carlo draw is used to estimate each agent's choice probability.* See simulation-based estimation for further details on the MSM estimator.

The idea behind the MSM estimator is quite general and can be applied in many other settings besides the multinomial probit model. McFadden's work helped to spawn a large literature on 'simulation estimation' that developed rapidly during the 1990s and resulted in computationally feasible estimators for a large new class of econometric models that were previously considered to be computationally infeasible. However, there are even better simulation estimators for the multinomial probit model, which generally outperform the MSM estimator in terms of having lower asymptotic variance and better finite sample performance, and which are easier to compute. One problem with the simple Monte Carlo estimator $\hat{P}(x_i, \theta)$ underlying the MSM estimator is that it is a discontinuous and 'locally flat' function of the parameters θ, and thus the MSM criterion function is difficult to optimize. Hajivassiliou and McFadden (1998) introduced the *method of simulated scores* (MSS) that is based on Monte Carlo methods for simulating the *scores* of the likelihood function for a multinomial probit model and a wide class of other *limited dependent variable models* such as Tobit and other types of censored regression models. (In the case of a discrete choice model, the score for the ith observation is $\partial/\partial\theta \log(P(d_i|x_i, \theta))$.) Because it simulates the score of the likelihood rather than using a method of moments criterion, the MSS estimator is more efficient than the MSM estimator. Also, the MSS is based on a *smooth simulator* (that is, a method of simulation that results in an estimation criterion that is a continuously differentiable function of the parameters θ), so the MSS estimator is much easier to compute than the MSM estimator. Based on numerous Monte Carlo studies and empirical applications, MSS (and a closely related simulated maximum likelihood estimator based on the Geweke–Hajivassiliou–Keane', GHK, smoother simulator) are now regarded as the estimation methods of choice for a wide class of econometric models with limited dependent variable that are commonly encountered in empirical applications (see simulation-based estimation for further details).

Mixed Logit Models

A mixed MNL model has choice probabilities of the form

$$P(d|x,\theta) = \int \left[\exp\left\{ \frac{u(x,d,\alpha)}{\sum_{d' \in D(x)} \exp(u(x,d',\alpha))} \right\} G(d\alpha|\theta). \right] \tag{3}$$

There are several possible random utility interpretations of the mixed logit model. One interpretation is that the α vector represents 'unobserved heterogeneity' in the preference parameters in the population, so the relevant choice probability is marginalized using the population distribution for the α parameters in the population, $G(\alpha|\theta)$. The other interpretation is that a is similar to vector ε, that is, it represents information that agents observe and which affects their choices (similar to e) but which is unobserved by the econometrician, except that the components of ε, ε (d) enter the utility function additively separably, whereas the variables a are allowed to enter in a non-additively separable fashion and the random vectors α and ε are statistically independent. It is easy to see that, under either interpretation, the

mixed logit model will not satisfy the IIA property, and thus is not subject to its undesirable implications. McFadden and Train proposed several alternative ways to estimate mixed logit models, including maximum simulated likelihood and MSM. In each case, Monte Carlo integration is used to approximate the integral in Eq. 3 with respect to $G(\alpha|\theta)$. Both of these estimators are smooth functions of the parameters θ, and both benefit from the computational tractability of the MNL while at the same time having the flexibility to approximate virtually any type of random utility model. The intuition behind McFadden and Train's approximation theorem is that a mixed logit model can be regarded as a certain type of *neural network* using the MNL model as the underlying 'squashing function'. Neural networks are known to have the ability to approximate arbitrary types of functions and enjoy certain optimality properties, that is, the number of parameters (that is, the dimension of the α vector) needed to approximate arbitrary choice probabilities grows only linearly in the number of included covariates x. (Other approximation methods, such as *series estimators* formed as tensor products of bases that are univariate functions of each of the components of x, require a much larger number of coefficients to provide an comparable approximation, and the number of such coefficients grows exponentially fast with the dimension of the x vector.)

Conclusion

This brief survey of McFadden's contributions to the discrete choice literature has revealed the immense practical benefits of his ability to link theory and econometrics, innovations that lead to a vast empirical literature and widespread applications of discrete choice models. Beginning with his initial discovery, that is, his demonstration that multinomial logit choice probabilities result from a random utility model with multivariate extreme value distributed unobservables, McFadden has made a series of fundamental contributions that have enabled researchers to circumvent the problematic implications of the IIA property of the MNL model, providing computationally tractable methods for estimating ever wider and more flexible classes of random utility and limited dependent-variable models in econometrics.

See Also

- ▶ Logit Models of Individual Choice
- ▶ Simulation-Based Estimation

Selected Works

1967. The evaluation of development programmes. *Review of Economic Studies* 34: 25–50.

1973. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, ed. P. Zarembka. New York: Academic Press.

1974. The measurement of urban travel demand. *Journal of Public Economics* 3: 303–328.

1974. (With R. Mantel, A. Mas-Colell, and M.K. Richter.) A characterization of community excess demand functions. *Journal of Economic Theory* 9: 361–374.

1976. The revealed preferences of a government bureaucracy: Empirical evidence. *Bell Journal of Economics and Management Science* 7: 55–72.

1976. (With M. Majumdar and T. Mitra.) On efficiency and Pareto optimality of competitive programs in closed multisector models. *Journal of Economic Theory* 13: 26–46.

1978. Cost, revenue, and profit functions. In *Production economics: A dual approach to theory and applications,* vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

1980. (With T. Mitra and M. Majumdar, M.) Pareto optimality and competitive equilibrium in infinite horizon economies. *Journal of Mathematical Economics* 7: 1–26.

1981. Econometric models of probabilistic choice. In *Structural analysis of discrete data with econometric applications,* ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

1981. (With C.F. Manski, ed.) *Structural analysis of discrete data with econometric applications.* Cambridge, MA: MIT Press.

1984. Econometric models of qualitative response models. In *Handbook of Econometrics*, vol. 2, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57: 995–1026.

1994. (With P. Ruud.) Estimation by simulation. *Review of Economics and Statistics* 76: 591–608.

2000. (With K. Train.) Mixed MNL models of discrete response. *Journal of Applied Econometrics* 15: 447–470.

2001. Autobiography. In *Les Prix Nobel. The Nobel Prizes 2000,* ed. T. Frängsmyr. Stockholm: Nobel Foundation.

Bibliography

Anderson, S.P., A. De Palma, and J.F. Thisse. 1992. *Discrete choice theory of product differentiation*. Cambridge: MIT Press.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov. 2005. Estimating static models of strategic interactions. Manuscript, University of Michigan.

Berkovec, J. 1985. New car sales and used car stocks: A model of the automobile market. *RAND Journal of Economics* 16: 195–214.

Berry, S., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890.

Block, H., and J. Marschak. 1960. Random orderings and stochastic theories of response. In *Contributions to probability and statistics*, ed. I. Olkin. Stanford: Stanford University Press.

Cosslett, S.R. 1981. Efficient estimation of discrete-choice models. In *Structural analysis of discrete data with econometric applications*, ed. C.F. Manski and D. McFadden. Cambridge, MA: MIT Press.

Dagsvik, J.K. 1983. Discrete dynamic choice: An extension of the choice models of Luce and Thurstone. *Journal of Mathematical Psychology* 27: 1–43.

Dagsvik, J.K. 1994. Discrete and continuous choice, max-stable processes and independence from irrelevant attributes. *Econometrica* 62: 1179–1205.

Dagsvik, J.K. 1995. How large is the class of generalized extreme value models? *Journal of Mathematical Psychology* 39: 90–98.

Debreu, G. 1960. Review of R.D. Luce '*individual choice behavior'*. *American Economic Review* 50: 186–188.

Dubin, J., and D. McFadden. 1984. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52: 345–362.

Eckstein, Z., and K. Wolpin. 1989. The specification and estimation of dynamic stochastic discrete choice models. *Journal of Human Resources* 24: 562–598.

Goldberg, P. 1995. Product differentiation and oligopoly in international markets: I the case of the U.S. automobile industry. *Econometrica* 63: 891–951.

Hajivassiliou, V., and D.L. McFadden. 1998. The method of simulated scores for the estimation of LDV models. *Econometrica* 66: 863–896.

Hajivassiliou, V.A., and P.A. Ruud. 1994. Classical estimation methods for LDV models using simulation. In *Handbook of econometrics*, ed. R.F. Engle and D.L. McFadden, vol. IV. Amsterdam: Elsevier.

Hannemann, M. 1984. Discrete/continuous models of consumer demand. *Econometrica* 52: 541–562.

Hausman, J., and D. Wise. 1978. A conditional probit model of qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46: 403–426.

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, MA: MIT Press.

Luce, R.D. 1959. *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Manski, C.F. 1985. Semiparametric estimation of discrete response: Asymptotics of the maximum score estimator. *Journal of Econometrics* 27: 303–333.

Manski, C.F. 2001. Dan McFadden and the econometric analysis of discrete choice. *Scandinavian Journal of Economics* 103: 217–229.

Pakes, A., and D. Pollard. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica* 57: 1027–1057.

Rust, J. 1994. Structural estimation of Markov decision processes. In *Handbook of Econometrics*, ed. R.F. Engle and D.L. McFadden, vol. 4. Amsterdam: Elsevier.

Thurstone, L.L. 1927. Psychophysical analysis. *American Journal of Psychology* 38: 368–389.

Meade, James Edward (1907–1995)

David Vines

Abstract

This article explains the part that Meade played in the creation of Keynes's *General Theory,* describes his work with Keynes during the Second World War in the creation of the IMF and the GATT, and summarizes the ideas in *The Theory of International Economic Policy* for which Meade was awarded the Nobel Prize

in 1977. It also sets out the role that Meade played in the construction of the inflation-targeting regime which became the centrepiece of British macroeconomic policymaking in the 1990s.

Keywords

Absorption approach to the balance of payments; Aggregate demand; Balance of payments; Brain drain; Chamberlin, E. H.; Compensation principles; Consumer surplus; Control theory; Demand management; Devaluation; Domestic distortions, theory of; Double-entry national accounts; Dutch Disease; Elasticities approach to the balance of payments; Exchange rate mechanism (ERM); Factor trade; Fiscal policy; Fixprice models; Fleming, J. M.; Foreign direct investment; Friedman, M.; Harrod, R. F.; Harrod–Domar growth model; Heckscher–Ohlin trade theory; Hicks, J. R.; Imperfect competition; Imperial preference; Income–expenditure theory; Incomes policy; Inflation; Inflation targeting; Institute of Fiscal Studies (UK); Interest rates; International policy coordination; International trade theory; IS–LM models; Johnson, H. G.; Kahn, R. F.; Kalecki, M.; Kay, J.; Keynes, J. M.; Keynesian Revolution; King, M.; Marshall, A.; Meade, J. E.; Monetarism; Monetary policy; Monetary theory of the balance of payments; Money multiplier; Money supply; Multinational corporations; Multiplier; Multivariable control methods; Mundell–Fleming model; National income accounting; New Keynesian macroeconomics; New welfare economics; Nominal-income targeting; Non-accelerating inflation rate of unemployment (NAIRU); Open-economy macroeconomics; Optimal savings rate; Optimum population; Phillips curve; Phillips, A. W. H.; Potential welfare; Profit sharing; Protection; Ramsey model; Regional free-trade areas; Robbins, L. C.; Robertson, D.; Robinson, E. A. G.; Robinson, J. V.; Saving and investment; Second best; Sidgwick, H.; Social security; Sraffa, P.; Stabilization bias; Sterilization; Sticky wages; Stocks and flows; Stolper–Samuelson theorem; Stone, J. R. N.; Supply side reform; Taylor rules; Taylor, J.; Terms of trade; Tinbergen, J.; Tobin's q; Trade creation and trade diversion; Unemployment; Wage fixing

JEL Classifications

B31

James Meade was one of the truly great economists of the 20th century. He was profoundly internationalist in his outlook, and was awarded the Nobel Memorial Prize in Economics in 1977, jointly with Bertil Ohlin, for *The Theory of International Economic Policy* (1951–5). But his contributions spanned the whole of the discipline. He made fundamental, and widely influential, contributions to economic theory, in both macroeconomics and microeconomics. More than this, his main concern was always with the part that economic analysis has to play in the solution of practical economic policy problems. As a result, he contributed to the theory of economic policy in a very wide range of subjects, including macroeconomic management, trade policy reform, public finance, economic growth, income distribution, wage determination, and population growth. He served actively in policymaking as an economist for the League of Nations, and in the Economic Section of the UK Cabinet Office during, and immediately after, the Second World War. In all that he did, Meade saw the role of an economist as helping to design a better society – both by the creation of good institutions of economic management and by the provision of appropriate incentives for private individuals.

In this article I concentrate on four main issues. I first explain the large part that Meade played in the creation of Keynes's *General Theory* in the 1930s. After this, I describe his work with Keynes during the Second World War in the creation of the International Monetary Fund and the GATT (which has since become the World Trade Organization, or WTO). I then turn to Meade's work on international economics at the London School of Economics (LSE), immediately after the war, for which he was awarded the Nobel Prize; I spend some time showing how this theoretical work was related to his earlier work on policy with Keynes.

Finally, I set out the role that Meade played, along with a group of young economists to which I belonged, in the construction of the inflation-targeting regime that became the centrepiece of British macroeconomic policymaking in the 1990s.

Activities Before the Second World War: Keynesian Macroeconomics

Meade was born on 23 June 1907 in Swanage, Dorset, and brought up in Bath. He went to school at Malvern College, and then won a scholarship in classics to Oriel College in Oxford. But like many others of his generation he was appalled by the problem of mass unemployment, which, as he said, caused 'poverty in the midst of plenty'. As a result he turned to the study of economics for the last two years of his university education. Meade gained greatly from studying classics, but, as a result of doing so, he had to teach himself the mathematics that he later used extensively.

Immediately upon graduating in 1930, Meade was elected to a fellowship at Hertford College, and appointed to a lectureship in economics at Oxford University. But in October 1930 his college first sent him to Cambridge for the academic year 1930/31, 'to learn my subject before I started to teach it. I had the greatest good fortune of being taken into Trinity College . . . by Dennis Robertson, to whose teaching that year I owe a deep debt of gratitude. At an early stage he told me that there was a young man in Kings called Richard Kahn whom I should get to know' (Meade 1983b, p. 263).

And so Meade spent a formative and creative year as a member of the 'Circus' which was gathered around Keynes. This group of people were debating Keynes' *Treatise on Money* (Keynes 1930) which had just been published, and included Joan and Austin Robinson, and Piero Sraffa, as well as Kahn. Meade enjoyed describing the 'workshop style' of the Circus meetings. Keynes took no part in the proceedings, but after each meeting Richard Kahn orally recounted to Keynes the subject matter of the discussions and the lines of argument.

> From the point of view of a humble mortal like myself Keynes seemed to play the role of God in a morality play; he dominated the play but rarely appeared on the stage. Kahn was the Messenger Angel who brought messages and problems from Keynes to the 'Circus' and went back to Heaven with the result of our deliberations. (Keynes 1971–88, vol. 13, p. 339; see also p. 338)

The casting of Keynes in this role was first suggested by Meade's wife Margaret in 1934 when they were staying for the weekend with Austin and Joan Robinson in Cambridge. That weekend, too, was dominated by messages from people who had just spoken with Keynes, though Keynes himself never appeared in person.

The Circus was discussing the failure of Keynes's *Treatise.* Keynes had expected that book to become his magnum opus. In it he set out the theoretical work which he had done since the end of the First World War, about the causes of the economic cycle, and about how this cycle should be managed on a national basis and internationally. There is much modern macroeconomics in the *Treatise*, and the international macroeconomics is particularly good: it is possible to find elements of the Swan diagram, of the Fleming-Mundell model, and even of the Dornbusch model (see Vines 2003). But the *Treatise* contains a fatal flaw. It aims to analyse the problem of the economic cycle, but the discussion rests upon a formal model in which the level of economic activity is exogenous. This mistaken model is nevertheless of interest because it contains the necessary clues about what Keynes needed to do next. In the *Treatise*, an increase in aggregate demand, caused – say – by an exogenous increase in investment, and not prevented by the central bank from having an overall effect on aggregate demand, would cause demand to increase relative to supply, which was assumed to be fixed, and so would cause a rise in the level of prices. This would redistribute income to profits, away from wages, because the level of the money wage is 'somewhat sticky' in the model. That would raise the overall level of savings, because the propensity to save is assumed to be higher for capitalists than for workers. A new equilibrium would be regained after the working out of a 'multiplier' process: one in which the

price level rises by the amount necessary to re-equilibrate leakages (the extra savings) with injections (the increased investment). This all makes sense, except if the model is meant to help in a discussion of booms and slumps in *output*, which it cannot do because output is exogenous. Sorting out this mess required Keynes to write the *General Theory of Employment, Interest and Money,* which took him until 1936.

By the time Meade arrived in Cambridge in October, Kahn had already drafted his famous article on the 'multiplier' (Kahn 1931). In this piece Kahn showed that, if output is endogenous, one can sum an infinite geometric series to show that the overall effect on output of an increase in investment is 'multiplied' because of the increases in consumption which happen as output increases. It appears that it was Meade, the young graduate student from Oxford, who showed how Kahn's multiplier analysis could be connected with Keynes's argument in the *Treatise.* There were two steps in this demonstration.

First, Meade showed, by summing the series of the effects of output on savings instead of the series of the effects of output on consumption, that movements in output would cause an increase in savings which would be equal to the original increase in investment. This idea, called 'Mr Meade's Relation' in Kahn's article, was written down in a note which was subsequently lost. It has become fundamental to our understanding of the multiplier process, and is explained in all basic macroeconomics textbooks. Meade (1993) explains the way in which his approach was complementary to that of Kahn. This approach was useful to the Circus, in that it showed how Kahn's multiplier process was a flex-output version of the fixed-output argument of the *Treatise* explained above.

Meade once described the second step of his demonstration to me as follows. 'I said the following to the other members of the Circus. "Haven't any of you read Marshall's *Principles of Economics?* In that book, in the short run, the economy lies on a short-run, upward-sloping, supply curve. But that curve adds an extra equation to the model. This means that – in comparison with the model in the *Treatise* – we can make *both* prices *and* output endogenous at the same time."' This second idea of Meade's is explained, with much less clarity, in Kahn's article. Once it was properly understood, it led the Circus to the view that it is primarily variations in the level of *output* which bring savings into line with investment, and so re-establish the conditions of macroeconomic equilibrium, rather than variations only in the level of *prices,* as had been supposed, unsatisfactorily, by Keynes in the *Treatise.* Kahn himself warmly acknowledged his debt to Meade, both in his article of 1931 and in his fascinating account of the period published in 1984 and called *The Making of Keynes' General Theory* (Kahn 1984).

Meade stated (Keynes1971–88, vol. 13, p. 342) that when he returned to Oxford in 1931 he took back with him in his head 'most of the essential ingredients of the subsequent system of the *General Theory*'. In his 'Simplified model of Mr. Keynes' system' (Meade 1937) Meade set out these 'essential ingredients' in a system of eight equations, which included those of the IS-LM model. It is known that Hicks saw a draft of this paper before he prepared his own celebrated article explaining the IS–LM system (Hicks 1937); indeed Hicks uses Meade's notation in his presentation (see Young 1987). Meade's 'simplified model' is more general than that of Hicks, because it includes the upward-sloping supply curve discussed above. It therefore enables one to see *much* more of what is going on in the *General Theory.* But Meade's article is very difficult to understand, because it takes so much for granted. One can read it carefully without ever seeing the point that Hicks was concerned to make about the relationship between Keynes and the 'classics', and it was Hicks who invented the famous diagram to explain this point. This appears to be the first of a number of occasions during Meade's career on which he would set out a fully specified piece of economic theory, only to find that someone else would extract a simple, essential, idea from what Meade had written, publish it, and become famous as a consequence.

Meade taught in Oxford until 1937. During this period he synthesized with great clarity the ideas of the Keynesian Revolution in *An Introduction to Economic Analysis and Policy* (1936), published

almost simultaneously with the *General Theory.* This was the first-ever economics textbook: until then books had been written as a means of expounding new ideas in economics. Many undergraduates in Cambridge at that time were confused by the turbulent debate concerning the Keynesian Revolution which swirled around them, and bemused by the associated misunderstandings, a number of which seemed to be deliberate. Subsequent oral tradition in Cambridge maintained that many of these people found Meade's book exceptionally helpful, since it cut straight through all of these difficulties.

Interestingly, the Keynesian model is expounded in Meade's book using an exogenous rate of interest. That is, Meade set out the 'Keynesian-cross' version of Keynes's model, rather than setting out the full IS–LM system. My own view of the reason for this is that Meade never really believed in the LM curve and so thought, like many of us now do, that the IS–LM system was a distraction. The reason that I say this is that in Meade (1937) he discusses the (realistic) possibility that banks might adjust the quantity of money, in the face of shocks to the economy, so as to keep the interest rate unchanged, unlike what happens in the IS–LM system (except in an extreme case). This view of his was in turn based on the analysis in his first published article (Meade 1934), in which he invented the money-multiplier – quite independently of the work in the United States begun by Phillips (1920) – and in which he carefully showed what banks would need to do in order to behave in this way. That Meade should have presented the Keynesian system in this way right back at the beginning, even although he fully understood the IS–LM system, has implications for understanding why – as discussed below – he proceeded the same way when he wrote *The Balance of Payments* in the late 1940s.

Economic Analysis and Policy also contains a discussion of longer-run growth, and presents an exposition of the Ramsey model of optimal growth. This was nearly ten years before Harrod and Domar invented what now looks like a very primitive version of growth theory, and long before the famous papers of Solow (1956) and Swan (1956), who invented a simplified version of the Ramsey model, in which the savings rate is *exogenous*. In two key pages, published 20 years before the papers by Swan and Solow, Meade explains how the optimal savings rate could be chosen *endogenously* in order to produce a welfare-maximizing growth process. These pages provide an astonishingly clear verbal exposition of the first order-conditions which must be satisfied if growth is to be optimal.

A second edition of *Economic Analysis and Policy* was published in 1937. This gave an exposition of the new ideas in imperfect competition, invented by Edwin Chamberlin and Joan Robinson, which were challenging Marshallian microeconomics in the 1930s (see Shackle 1967). These ideas only really bore fruit in mainstream microeconomics, and in macroeconomics, in the 1970s and 1980s, after the rise of game theory. But they had a more or less immediate effect on Meade's work in macroeconomics, as I will show below.

Economic Analysis and Policy also includes a section on problems of international order and disorder, which shows that, already as a young man, Meade (unlike many in Britain and America at this time) was thinking about the macroeconomic problems of the world, as distinct from those of an individual nation. In this part of the book, Meade expands on the ideas on international macroeconomics, which were already to be found in Keynes's *Treatise,* as I have described above. (Notably, Joan Robinson 1937, and Roy Harrod 1933, were also busy doing the same thing.) Meade's volume ends with a prescient chapter on the economic causes of war. It is chilling to re-read this chapter, written three or four years before the outbreak of the Second World War. One also realizes that many of the problems connected with internationally ill-coordinated macroeconomic policies, which emerged in the world economy in the 1980s, and which have re-emerged at the beginning of the new millennium, are very like those which Meade wrote about nearly 75 years ago.

Throughout his time at Oxford, Meade was actively involved with the group of Fabian socialist intellectuals who were helping the British Labour Party to recover its sense of purpose, after the

disastrous collapse of the Labour Government in 1931. Meade contributed to discussions across the same wide range of macroeconomic, microeconomic and international issues that he had treated in *Economic Analysis and Policy*, and he was most influential in his advocacy of expansionary Keynesian policies (Durbin 1985, see especially pp. 136–44, 194–8, 211–12 and 220).

Meade elaborated on this last theme in *Consumers' Credits and Unemployment* (1938). In this work, he proposes Keynesian demand management in the form of *automatic*, countercyclical variations in taxation to stabilize macroeconomic fluctuations. This book foreshadows both the Full Employment White Paper of 1944, and the nominal-income targeting project on which he worked for ten years from 1978, both of which are discussed in more detail below. *Consumers' Credits and Unemployment* is perhaps the earliest official published advocacy of fine-tuned Keynesian policies.

In 1937 Meade went with his wife and young family to Geneva, where he was to stay for three years, as an economist for the League of Nations. Meade often spoke with admiration of the remarkable band that were assembled there, which included Tinbergen, Koopmans, Haberler, Nurkse, and Marcus Fleming. His job was to prepare, more or less single-handed, the *World Economic Survey* (the forerunner of the present IMF *World Economic Outlook*) and he transformed this publication. Keynes was much influenced by Meade's work in Geneva. In the *General Theory,* Keynes had made use of an upward-sloping short-run supply curve, as described above in my account of the discussions of the Circus, which relies on goods being produced in a manner subject to diminishing returns, and being supplied under competitive conditions. Keynes's important article of 1939 makes extensive reference to Meade's work, and then goes on to argue that the quantity produced in an economy is determined by demand, even if there are constant marginal costs. But this was inconsistent with the competitive analysis that Keynes had utilized in the *General Theory*, which requires the assumption of increasing marginal costs. This article by Keynes was exceptionally difficult to understand. It was extensively discussed in Cambridge in the 1970s, when people were comparing Keynes's macroeconomics with that of Kalecki, who had carefully evaded this difficulty by assuming 'markup' pricing. The confusion was resolved only by the arrival of the classic Dixit and Stiglitz (1977) paper. That paper brought Chamberlin's ideas about imperfect competition into macroeconomics, suggesting a setup in which each individual profit-maximizing producer faces a downward-sloping demand curve and sets prices above marginal costs, but in which the existence of free entry prevents the emergence of monopoly profits. I know, from working with Meade from the late 1970s onwards, that the standard macroeconomic model which he used daily, as part of his mental equipment, had this feature, and I believe that, unlike many others, this had been true for him since the mid-1930s, when he wrote the material on imperfect competition in *Economic Analysis and Policy*, which I described above. It is probable that this framework influenced his empirical work in Geneva, and thus influenced Keynes's article of 1939.

The war finally caused the Meades to leave Geneva for London in 1940, with three young children, the smallest a three-week old infant. They set out in a small car for one of the Channel ports, not knowing that at this very time the Germans had broken through at Sedan. After an increasingly desperate journey the family ended up in Nantes as refugees, and finally crossed the Channel in an RAF 'transport ship' – a converted tramp steamer – at the very time of the Dunkirk evacuation.

The War Years, 1940–45: Building the Post-War World Order

On his return to Britain, Meade was brought into the Economic Section of the Cabinet Office. There was a grim feeling of impotence amongst the economists, who wished to do something for the war effort. Keynes's *How to Pay for the War* (Keynes 1940) had just been published. Meade therefore set to work on a set of national accounts, so that, at the very least, those making war policy might be able

to attach some numbers to Keynes's ideas. Richard Stone who had only recently graduated in economics from Gonville and Caius College, Cambridge, was brought in to help with this work. Together they produced what is probably the first full logical structure of 'double-entry' national accounts (Meade and Stone 1941, 1944). Meade recalled how in the statistical work he quickly became Stone's research assistant – and so began Stone's work on national income accounting that eventually led to another Nobel Prize.

During his subsequent time at the Economic Section, Meade worked in three crucial areas. He was to become Director of the Economic Section from 1945 to 1947.

First, Meade was involved in the planning for post-war international monetary arrangements. He participated in the initial excited responses in Whitehall to Keynes's 'Clearing Union' plan for a new post-war international monetary system (Keynes 1971–88, vol. 25, pp. 41–67; see also van Dormael 1978). He became a member of the British delegation to Washington in September 1943 which discussed these issues with Harry Dexter White and others (Keynes 1971–88, vol. 25, pp. 338 ff.). And he took part in the subsequent British deliberations, leading up to the Bretton Woods conference in 1944 at which the International Monetary Fund was established.

I have described the analytical content of these negotiations in some detail in Vines (2003), drawing on the wonderful historical account by Skidelsky (2000), and on the papers of Keynes and Meade. Keynes's policy objectives were to create a post-war global system in which full-employment policies could be adopted by the Allied nations, and in which such full-employment policies could be reconciled with the requirement that their trade balances not get too far out of line. Keynes's initial response to this problem was a highly illiberal one: balance-of-payments restrictions should be the mechanism which re-equilibrated exports with imports after any negative external shock to a country (through tariffs, quotas, and 'managed' trade). He was persuaded away from this view by an 'outstandingly able group of economists' (Williamson 1983a, p. 91) which included Meade and also Marcus Fleming, Roy Harrod, Lionel Robbins, and Dennis Robertson. This group managed to convince Keynes that exchange rate devaluation should be the adjustment device. Keynes vacillated literally for years on the issue, deliberately suspending judgement, drawing forth from this talented group an extraordinary collection of papers, particularly on the tariffs-versus-devaluation issue (Keynes 1971–88, vol. 26, ch. 2 and pp. 239–327). James Meade once told me that, one day in a particularly tedious meeting at the Board of Trade in 1944, Keynes scribbled a note to him to the effect that he (Keynes) was, at last, intellectually converted to a regime in which external adjustment would be achieved by exchange-rate change.

Second, when Keynes produced his Clearing Union plan, Meade quickly produced a project for a 'Commercial Union' as a companion piece. It was on the basis of this document that the debate in Whitehall on post-war commercial policy (concerning such sensitive issues as imperial preference and the use of import restrictions on balance-of-payments grounds) took place. Meade devoted much time to drafting and redrafting these ideas and, as he said, 'helping to get them through Whitehall'. And it was to promote these ideas that he was a member of the September 1943 delegation to Washington (mentioned above), and he was subsequently a member of the British delegation to the international conferences in London in 1946 and in Geneva in 1947 which worked on a charter for a proposed International Trade Organization (ITO). Although in the end the ITO proved to be unacceptable to the United States, the Geneva conference resulted in a General Agreement on Tariffs and Trade (GATT) which took on many of the projected functions of the ITO (see Keynes 1971–88, vol. 26, ch. 2). And the GATT was eventually turned into the World Trade Organization (WTO) in 1994.

These international discussions laid down, amongst other things, the conditions under which nations should be permitted to form regional free-trade-areas, in which discriminatory regional preference is allowed to overrule the most-favoured-nation rule for international trade which lies at the centre of the WTO, and which lay at the centre of the GATT. Those discussions duly

led to Article 24 of the GATT (and to a similar provisions in the international agreements which underpin the WTO). The technical discussions on this Article were particularly difficult, since the relevant theory by Viner and Meade, on 'trade creation' and 'trade diversion', had not yet been invented. (This theory is discussed in section "The LSE, 1947–58: International Economics" below.) The discussions also contained much which was non-technical, which dealt more fundamentally with the nature of the international trading system. On one occasion, Meade told me, he could not understand why a senior US official – I believe that it was Dean Acheson – was speaking up so strongly against imperial preference, and yet so much in favour of Britain joining up with European nations, a joining-up which has, in due course, led to the European Common Market, and ultimately to the creation of the European Union. 'I have relatives who are farmers in New Zealand', said Meade, 'who sell their lamb to Britain. They a natural part of the British economic system. Why should we not have an Imperial Free Trade Area which includes them? This would be just like your setup in the US, in which you have a free trade area which includes all 50 of your states?'. 'But there is a lot of water between Britain and New Zealand', replied Acheson. 'There is also quite a lot of water between Britain and France', replied Meade.

If we take these two activities together, it is clear that Meade was one of the architects, on both the monetary side and the trade side, of the liberal world economic regime which sustained the long post-war boom in the Western world from 1945 to 1973. Meade always believed that *both* pieces of this regime stand or fall together. Free trade would – he thought – be resisted if there were severe global macroeconomic imbalances. (This point became clear once again in the mid-1980s, and it is becoming even more clear in the mid-2000s. But conversely, if there is not free trade then macroeconomic order will be difficult to maintain, since devaluation will tend to be much less effective at adjusting trade imbalances. Meade summarized this point clearly, in a diary entry which he made on 31 December 1944 (Meade 1988–90, vol. 4, p. 22.) He emphasized 'the need for flexible exchange rates to adjust balance of payments [to avoid pushing the burden of adjustment onto] rigid trade controls . . . in a world in which internal wage levels were not easily reduced. [But such adjustment might be] more easily acceptable if it was preceded by an international agreement to lower trade barriers, since in that case smaller movements in exchange rates would be required'. This belief – that macro management and micro liberalism should go together – had already informed his work in the 1930s. It would form the central organizing principle for the work that Meade did at the LSE on international economics, which I discuss immediately below. As noted in the introduction and conclusion to this article, it recurs again and again throughout his work.

At the meeting at the Board of Trade in 1944, to which I referred above, Keynes followed up his scribble with a sketch, on the back of an envelope, of the desired features of the whole international system that he and his colleagues were trying to build (see Vines 2003). This sketch went something like the following.

Objective	Instrument (s)	Responsible authority
Full employment	Demand management (mainly fiscal)	National governments
Balance of payments adjustment	Pegged but adjustable exchange rates	International Monetary Fund
Promotion of international trade	Tariff reductions etc.	International Trade Organisation
Economic development	Official international lending	World Bank

Keynes was aware that this plan would need to work, not just for individual countries, but for the global system as a whole. (It would have been surprising if someone who had invented macroeconomics did not take such an overall, systemic view.) I discuss in some detail in Vines (2003) how Keynes feared that difficulties in the balance-of-payments adjustment process might impose, on deficit countries, an obligation to deflate demand below full employment, something which might

not be matched by symmetrical over-expansion by surplus countries, and might thereby create pressures towards global deflation. I also describe how Keynes differed in this view from Harry Dexter White, his US counterpart in the Washington negotiations, who feared an outcome in which the International Monetary Fund would be so expansive with liquidity that there would be a great post-war inflation, worldwide. In that article I claim that, during these discussions, Keynes' negotiating strategy in pursuit of a balanced global outcome was underpinned by a significant theoretical understanding of what was going on. In particular I maintain that (*a*) Keynes took from his *Treatise* something akin to an IS–LM–BP model (without the flaw in the analysis of the *Treatise*, which had been fixed up by the invention of the multiplier and the publication of the *General Theory*), and (*b*) Keynes, as he negotiated, was using something akin to a two-country version of that model to understand what was being discussed. These two claims of mine are vital for a proper understanding of the work that Meade did at the LSE on international economics, which I discuss below.

I will be brief about Meade's third activity while he was in the Economic Section during the war, although it was important. Meade's paper 'Internal Measures for the Prevention of General Unemployment', dated 8 July 1941, reached the Inter-Departmental Committee on Post-War Internal Economic Problems in November, and as Skidelsky (2000, p. 270) says 'never quite lost its place as front-runner in the development of post-war employment policy'. This was, in effect, the first draft of what finally became the Full Employment White Paper, published as an official paper with the title of *Employment Policy* (Minister of Reconstruction 1944), which laid the basis for a transformed macroeconomic management within the United Kingdom after the war. In the drafting of this document, there were long discussions between Meade and Keynes on the possibility and desirability of automatic fiscal fine-tuning (see Keynes 1971–88, vol. 27, pp. 207–19 and 308–79; Wilson 1982). Meade advocated counter-cyclical variations in social security contributions; this proposal featured in the final White Paper and was endorsed by Keynes. That idea would remain more or less an article of faith for Meade, and underpinned his work in inflation targeting which I describe in section "Retirement, 1969–95".

The LSE, 1947–58: International Economics

Meade became Professor of Commerce (with special reference to international trade) at the LSE in 1947, where he was to stay for ten years, and where his great work on international economics was done.

It had been Meade's intention to begin his time back at a university by rewriting his textbook *Economic Analysis and Policy.* But this was not to happen. Someone once observed to me how different the teaching of our subject might have been if Meade had actually rewritten his book, rather than leaving the field open for Samuelson's great *Principles* book (Samuelson 1948), which was not published until 12 years after Meade's book had first appeared. In his Nobel Prize autobiography Meade (1977a) explained why this did not happen.

> I realised that it might be necessary to [rewrite the book] in more than one volume. So, as I was appointed at the LSE to teach international economics, I started on *The Theory of International Economic Policy.* It grew into my two books, *The Balance of Payments,* and *Trade and Welfare,* with their two mathematical appendices. . .. These books took up practically the whole of my ten years at the LSE; but even so they did not cover the whole of the international problem. . .. My original project was over-ambitious; but the part which I did manage to cover was sufficient, eventually, to gain for me the Nobel award. (Meade 1977a)

It is characteristic of Meade's modesty that he should describe the work for which he received the Nobel Prize as an attempt to rewrite a textbook.

The Balance of Payments

In his introduction the *The Balance of Payments* (Meade 1951–5), he had, equally modestly, described it as a book which 'does not claim to make any significant contribution of original work

in the fundamentals of pure economic analysis' (p. vii). This is something which turns out not to be true.

Meade also said of his book that it is one which has an 'indebtedness to the ideas of Lord Keynes [which] is too obvious to need any emphasis' (p. ix). Many people have said to me that they think that this remark is there because the book contains lots of 'multiplier Keynesianism', of a kind derived from the *General Theory,* which was still new and exciting in the 1950s. If that reading is correct, the generous acknowledgement of Keynes's contributions would not be particularly significant. But I believe the remark meant something rather different and rather more interesting. On more than one occasion Meade said to me that all he had done in this book was to write down what he learned from his work with Keynes during the war, about how to understand the international position of the British economy, and about how the world economy should be managed. That is a much more thought-provoking connection to acknowledge. (He did also admit that he had added quite a lot of algebra in the appendix.)

Volume I of the *Theory of International Economic Policy* (Meade 1951–5) was entitled *The Balance of Payments.* There were three new features of this book. First, at the level of technical analysis, it integrated income effects and price effects so as to study the balance of payments in a general-equilibrium framework. In doing so, it extended the theory of the balance of payments beyond its traditional identification with the current account to so as to consider the overall balance by including international capital movements. Second, it had a policy orientation, focusing on two instruments (exchange rate adjustment and domestic demand management) and two targets (internal balance – that is, full employment – and external balance – that is, a satisfactory overall balance of payments position). Third, Meade carried out his tasks in this book using a two-country model rather than merely developing the analysis for a single open economy.

At the level of technical analysis, investigations of the effects of exchange rate change had previously been separated from investigations of Keynesian income–expenditure theory. The former was normally based on the assumption of constant incomes, and carried out in terms of Marshallian partial equilibrium concepts, using the elasticities approach. (For a few key exceptions, see Robinson 1937; Harrod 1933; Laursen and Metzler 1950; and Harberger 1950.) The latter was normally carried out using fixprice models, which led to the 'absorption' approach to the balance of payments, published by Alexander in the same year as Meade's book (Alexander 1952). The formal integration of the elasticities approach and the absorption approach to balance-of-payments theory, which Meade achieved, was very important.

At the level of the theory of economic policy, Meade's basic idea – that, if internal and external balance are to be attained simultaneously, then two policy instruments are needed (exchange rate adjustment and the management of domestic demand) – was not a new one to him. He would have been familiar with this idea from his work with Keynes at the beginning of the war on Keynes' book *How to Pay for the War,* and also from his work with Keynes on Britain's financial crisis at the end of the war. (See Vines 2003, for a detailed discussion of this claim.) Furthermore, as noted at the beginning of this article, many of the necessary components of this idea are already to be found in the *Treatise*, published more than 20 years earlier; and many of them are also to be found in *Economic Analysis and Policy,* published 15 years earlier, and in the work of Robinson and Harrod referred to above.

Indeed, this idea now seems deeply obvious to all of us. But that is only because we know the Swan diagram, which collapses all of the complex analysis by Meade into just one diagram (Swan 1963), just like Hicks had done with the IS–LM system. At the time, Meade's idea was absolutely revolutionary. In reminding ourselves of this fact, we should not forget that Tinbergen was awarded the Nobel Prize in 1969 for stating a more general, but equally obvious, idea – that to achieve n targets simultaneously one (normally) needs n instruments. (Tinbergen's analysis was developed simultaneously to, and independently of, Meade's book.) And it took a *very* long time for Meade's idea to be learned. For many years after

the Second World War in the UK, full employment policies appear to have been carried out without sufficient regard for their effects on the balance of payments, and they often needed to be reversed at times of balance-of-payments crisis. Also, to take another example, many policymakers still continue to forget that if a devaluation is to improve a current-account deficit then it must be accompanied by a reduction in domestic absorption relative to domestic output, so as to release the resources needed to improve the trade account.

All of what I have said so far is about open-economy macroeconomics. We should also notice the third important feature of Meade's book which I have mentioned above – that it develops everything for a two-country world, and discusses *global* macroeconomics, not just *open-economy* macroeconomics. You might think that this would be the obvious way to proceed. After all, any treatment of *trade* theory is normally done this way, by analysing trade in a two-country world, and this is what Meade himself would do in Volume II of the *Theory of International Economic Policy*, published a few years later. Furthermore, all of us have now lived through the 1980s, in which we studied the effects of Reaganomics on Europe, something which clearly required a *two*-country model. (We are all at present trying to understand the interrelationships between the United States, East Asia and Europe, which seems to need a *three*-country model.) But nobody had ever done global macroeconomics before Meade wrote his book. As I note in Vines (2003), even Keynes, when writing down of the key components of the necessary theory in the *Treatise* in 1930, wrote about nearly everything for a single open economy rather than for a global system. But in Vines (2003) I also develop the argument, described at the end of section "The War Years, 1940–45: Building the Post-War World Order" above, that Keynes worked out, informally, aspects of the needed two-country model when he was negotiating with Harry Dexter White about the establishment of the IMF. It is my belief that Meade had seen, when working on these negotiations with Keynes, that such a model was necessary for a systemic discussion of global, policy-related, questions. This is my view of why he set out his analysis in this way, even although doing this made his book *much* harder to read.

Harry Johnson made two important criticisms of Meade's book at the level of technical analysis. The first, developed in Johnson's long review of the book (Johnson 1951), concerned the treatment of saving in the model. Meade assumes that the amount of real saving coming from a given real income is independent of the terms of trade. Laursen and Metzler (1950) and Harberger (1950) had already shown how to avoid this mistake; many practitioners of open economy macroeconomics still forget how hard it is to defend what Meade assumes.

Johnson's second criticism, made in the paper which Johnson published at the time the Meade received the Nobel Prize (Johnson 1978), leads in a valuable direction. What Johnson said was that Meade did not succeed in fully integrating real and monetary theory in his book. What he meant by this is that Meade assumed a flexible money supply policy designed to maintain a given exogenous rate of interest, with monetary policy changes being expressed in terms of (exogenous) interest rate changes. In the IS-LM-BP model subsequently developed by Fleming (1962) and Mundell (1962), the interest rate instead becomes endogenous, so as to ensure that the economy lies on a given LM curve. Under fixed exchange rates, the LM curve moves around because of the endogeneity of the money supply, unless monetary sterilization is possible, in a way which was analysed in the monetary theory of the balance of payments (which led to fame for Harry Johnson). Under floating exchange rates the money supply is held constant, and the interest rate and the exchange rate together become jointly endogenous, along with output. It seems odd that Meade did not think to make the interest rate endogenous by introducing an LM curve into his model, since this is exactly what he had done, more than 15 years earlier, when he had explained Keynes's *General Theory* to the world. Had Meade done this, surely he would have instantly invented the Fleming–Mundell model. My suggestion of the reason why that didn't happen is related to my view, stated earlier, that Meade never really

believed in the LM curve. In his subsequent work, which I discuss below, work that was contemporaneous with that of John Taylor, Meade allowed for the endogeneity of the interest rate without having to make the ridiculous assumption of a fixed money supply – essentially by supposing that the interest rate would follow something like a Taylor rule. One can easily build a Fleming-Mundell-like model with a Taylor rule in it, instead of an LM curve. I believe that, although Meade did not like the way that the interest rate was made endogenous in the LM curve, at the time he was writing *The Balance of Payments* he could not yet see how to replace the LM curve by a policy–behaviour relationship like the Taylor rule. This is why, I think, it was not possible for him to take the next step and construct something akin to the Fleming–Mundell model.

Trade and Welfare

The second volume of the *Theory of International Economic Policy* was titled *Trade and Welfare.* In this, Meade presented a systematic analysis of neoclassical trade theory, essentially the theory of Heckscher and Ohlin, with the latter of whom he shared the Nobel Prize. But he combined this with an analysis of trade in factors – both capital and labour. He discusses policy in this book – the issue of protection versus free trade – but in relation to the movement of *both* goods and factors of production. Meade's inclusion of international factor movements, in the main corpus of his theory of international trade, was innovative, and chimed with growing concerns, at the time, and since, about the 'brain drain', foreign direct investment, and the multinational corporation. Surprisingly, very few expositors of trade theory have followed Meade in explaining trade theory in this way, so that these subjects are more normally studied in isolation. Perhaps, again, it is because Meade's integrated analysis makes for such difficult reading.

The book made a number of important innovations at the level of technical analysis, whose influence in economic theory went far beyond the study of international phenomena.

First, Meade introduced a new method for measuring small changes in welfare, which was a generalization of Marshallian consumer surplus, with its attendant limitations. And he then went on to present a whole new approach to welfare economics, defining overall welfare as an appropriately weighted sum of individual welfares. Johnson (1978) describes it as a brilliant feat of imagination for Meade to see that he could take over this approach from Fleming (1951) and then rework it into a powerful general technique for welfare analysis of practical policy problems. Doing this enabled Meade to escape from the nihilism of the new welfare economics, which worked in terms of 'potential welfare' and the 'compensation principles', but which made it difficult to say anything practical at all about the welfare effects of economic policy changes. Nearly all of us now do welfare economics in the manner pioneered by Meade.

Second, Meade invented the theory of domestic distortions in order to show that a move towards free trade may not be welfare-improving if there are already distortions elsewhere in the economy. This idea was later carried forward by Bhagwati and Ramaswami (1963) and Johnson (1965). Meade invented the theory of the second best in his discussion of these ideas (inventing the technical term 'second best' as he did so), and explored many of its implications. As Corden (1996a) says, it is hard to see how something which now seems so obvious needed to be invented. Jacob Viner, in a book on customs unions (Viner 1950), had already established the distinction between trade creation and trade diversion in the creation of free trade areas. This also seems totally obvious to us now, but it really only became obvious after Meade published the *Theory of Customs Unions* (1955b), which clarified and extended Viner's distinction, and located it within his general theory of the second best.

Finally, it is important to add that the *Trade and Economic Welfare* includes a discussion of the meaning of optimum population and of optimum savings and of the relationship between these two concepts. This discussion, too, broke new ground.

Phillips

While at the LSE, Meade also did something else which was stunningly important: he brought Bill

Phillips into economics. Meade once said to me that Phillips was the closest to genius of anyone that he had ever known. Phillips's really important work in economics was on the use of control theory for macroeconomic stabilization purposes, rather than in estimating the 'Phillips curve' (for which he is so famous, but which he did in a rush in a few weeks, just before leaving London go on sabbatical leave).

Phillips had trained before the war as an electrical engineer (having previously left school without any formal qualifications), and immediately after the war he had graduated from the LSE with a third-class honours degree in sociology. One day, soon after receiving this unremarkable qualification, Phillips explained to Meade that he wished to build a strange 'water-machine' model of a macroeconomic system. Meade listened patiently because 'the pipes seemed to have the right labels', and so encouraged Phillips to build the machine, offering Phillips the inducement that he could demonstrate it at Lionel Robbins's prestigious seminar for graduate students. The machine was duly built, and it is described in Phillips (1950). A brilliant performance followed at the Robbins seminar, in front of most of the London economics professoriat, who had got word of what was coming. In the course of that seminar, said Meade, Phillips gave the best exposition that anyone present had ever heard of the Keynes-versus-Robertson debate, about whether the rate of interest was determined by liquidity preference or by the supply of, and demand for, loanable funds. This, said Phillips, was an argument about stocks versus flows; he then illustrated his claim by displaying the effects of water sitting in tanks, on the one hand, and water flowing through pipes, on the other. Phillips was duly instructed to write up his machine in a Ph.D. thesis, and John Hicks, who was by then Drummond Professor of Political Economy in Oxford, was asked to examine the thesis so as to ensure that somebody with a third-class degree in sociology could be given a Ph.D. in economics with a clear conscience. Phillips was then promptly brought on to the staff, and became one of the professors in the department within a few years. In Vines (1996) I give a detailed account of how the Phillips machine works. In particular I describe the stock-flow intuition which it provides, which is almost impossible to obtain any other way than by looking at this machine in action, and which certainly cannot be obtained from modern computer simulation models. As I describe below, Meade was closely involved with the use of the machine.

Work on his machine led Phillips to write his classic article on the use of control theory to help stabilize an economy (Phillips 1954). This paper argued that a feedback policy can have destabilizing effects if the instrument of policy responds too strongly to a disturbance to the target of policy, and there is a lag in the effect of the instrument on the target. In a subsequent paper, Phillips concluded on a cautious note: 'the problem of economic stabilisation is, even in principle, a very intricate one, and . . . a much more thorough investigation of both theoretical principles and empirical relationships would be needed before detailed policy recommendations could be justified' (Phillips 1957, p. 275). Meade was involved with the preparation of both of these papers, and he agreed with their conclusions.

Milton Friedman came to hold similar views on the potentially destabilizing effects of macroeconomic policy, at a very similar time (Friedman 1953). He went on to declare that active macroeconomic policymaking is too difficult to do properly and, worse still, too dangerous. Friedman's response to this problem was to set off in pursuit of his holy grail: a non-interventionist macroeconomic policy.

Meade's and Phillips's response to this problem was rather different. Phillips thought that it would be possible to do good macroeconomic policy, but only if the policy was carefully designed. Indeed he ended his 1957 paper on an optimistic note. He called for the use of multivariable control methods, to regulate multiple objectives in an economy in the face of multiple disturbances, and he noted that methods for doing this were just, in the late 1950s, becoming available. He also called for the econometric estimation of the parameters of the econometric model which would be necessary for the study of such regulation. Meade said to me on more than

one occasion that he regarded his own last big project, carried out more than 20 years later, and described in section "Retirement, 1969–95" below, as a response to Phillips's call to action.

Other Activities

At the LSE Meade acquired a further generation of very able young disciples, drawn from many countries, who included Max Corden, Richard Lipsey, Robert Mundell and Harry Johnson, the last of these 'at one remove' (Johnson 1978, p. 66). Meade had persuaded Phillips to build two of his water machines, joined together by an ingenious model of a foreign-exchange market. Peter Kenen (now retired from Princeton) vividly remembers a graduate student seminar in which he was asked to run fiscal and monetary policy for the United States on one of these machines. At the same time on the other machine Richard Cooper (now retired from Harvard) was required to run fiscal and monetary policy for Europe. They made the world develop unstable cycles – and spilt a lot of water (Vines 1996). By such means did that generation of students learn about the need for an international coordination of macroeconomic policies, 25 years before the subject became fashionable.

During this time Meade also went on sabbatical leave to Australia. With Eric Russell of Adelaide he wrote a short theoretical analysis of the effects of the Korean War boom on the Australian economy, via its effects in raising the world price of wool, which was – at that time – a major export commodity for Australia (Meade and Russell 1957). This article is one of the most profound pieces ever written about that economy. (Harcourt 1982, ch. 21, describes how it came to be written: Meade became the expositor of Russell's perceptions, which then existed only in note form.) The authors first explain the 'Stolper–Samuelson' theorem concerning the effects of protection on income distribution. Their exposition is different from the one given by Stolper and Samuelson, and much more like that to be found in the original source of that theorem – the Brigden Report of 1929 on the Australia tariff (Brigden et al. 1929). This is because it discusses the effects of protection on income distribution in an Australian 'dependent-economy' model, in which there are non-traded goods as well as traded goods. (See Vines 1994.) Meade and Russell then use this model to examine what has subsequently become called the 'Dutch Disease'. They show how an export boom can, by raising wages, give rise to cost pressures for the protected sector, which can cause it to contract, even at a time of general boom. Their paper directly influenced the subsequent discussion of this problem, first in Australia in the 1970s (see Gregory 1976), and then worldwide in the 1980s (see Corden 1984).This 'problem' has returned in a big way in the early 21st century, with the high prices of primary commodities, worldwide.

Cambridge, 1957–69: Growth Theory

Meade became Professor of Political Economy at Cambridge in 1957, when he succeeded his teacher, Dennis Robertson. When Meade moved to Cambridge, growth theory was in the air. His useful book, *A Neo-Classical Theory of Economic Growth* (1961b), 'brings the subject within the range of the undergraduate student, covers a number of aspects (such as the presence of the fixed factor land) usually omitted in more high powered mathematical treatments, and presents in detail the mathematics of a two-sector growth model' (Johnson 1978, p. 79). He also made advanced contributions to growth theory (1965, with Frank Hahn, and 1966). But his lasting contribution in this area is his essay *Efficiency, Equality, and the Ownership of Property* (1964). This 'provides a very suggestive account of the forces underlying the accumulation of capital and the relationship between earned and unearned income' and 'stimulated much of the revival of interest in this subject, at least in the United Kingdom' (Corden and Atkinson 1979, p. 530). Meade regarded this as, in many ways, his best book, because it puts together into a single synoptic framework his views on economic growth, on the microeconomic role of the price mechanism, on the size and the genetic composition of the population, and on the distributional implications of property ownership. He analysed further the interplay of these last factors

in his Keynes Lecture on 'The Inheritance of Inequalities: Some Biological, Demographic, Social, and Economic Factors' (Meade 1973c).

In 1960 Meade visited Mauritius and contributed to a report to the Governor, applying for the first time his ideas on growth theory and on population policy to the problems of a less developed country (Meade 1961a). His prediction for Mauritius of Malthusian stagnation turned out to be spectacularly wrong, in interesting ways.

In 1973 Meade also began in Cambridge a grand scheme of work entitled *The Principles of Political Economy.* The purpose of this series of books was 'to bring the best of modern theory within the range of an intelligent and educated adult, the volumes being intended to tackle successively departures from the assumptions of a model of perfect static general equilibrium' (Johnson 1978, p. 79).

Retirement, 1969–95

In 1969 Meade took early retirement, five years before the statutory retiring age. As Atkinson and Weale (2000) say, '[a]lways the most gentle and courteous of men, he had found extremely depressing the quarrels between those labelled "post Keynesian" and those in the Faculty who researched the mainstream of Economics'. But he did not stop working; indeed the next quarter century was to be one of his most productive.

Meade initially worked on the *Principles of Political Economy* but subsequently, perhaps sensing that this enterprise did not provide the best outlet for his unflagging energy, he turned to other schemes. The *Intelligent Radical's Guide to Economic Policy* (1975a) had 'wide influence in Britain, particularly on debates about economic planning' (Corden, and Atkinson 1979, p. 530). In it Meade returned to a theme set forth in his *Planning and the Price Mechanism* (1948a) which I summarize in my concluding section below.

An Expenditure Tax

Meade's first big activity in retirement was to chair a committee which was established by the Institute of Fiscal Studies, to look into the structure of the UK tax system and to advise on how it might be simplified. The report, entitled *The Structure and Reform of Direct Taxation* (1978a), is a monumental study of British personal taxation. As Atkinson and Weale (2000) say,

> The Committee observed that the tax system at the time was a mixture of taxes on income and taxes on expenditure, and concluded that it should be more desirable that tax should be levied on one or the other, all but one of the Committee favouring a shift towards an expenditure tax. In the 20 years since the report was written, exemptions for saving have appeared in the form of TESSAs, PEPs and ISAs, and the shift to indirect taxation has been a move towards a tax on expenditure. In this respect, the Report was influential, but its lasting value lies in the outstandingly high quality of the analysis.

Meade was fortunate in having as assistants for that committee three additional able young scholars, John Flemming, John Kay and Mervyn King, who all subsequently achieved distinction in various aspects of public life.

A Return to the Theory of Macroeconomic Policy

In 1977 Meade returned to the great questions of national macroeconomic management, at the age of 70, when most people might have felt ready for a holiday. His work began with his Nobel Prize lecture entitled 'The Meaning of Internal Balance' (Meade 1978b). It has been explained above how, in the *Balance of Payments* (1951) – one of the volumes for which Meade received the prize – he talked about the problems of reconciling internal balance (full employment) with external balance (a satisfactory overall balance of payments position). In his Nobel Prize lecture, Meade returned to question this framework, arguing that the concept of 'internal balance can now no longer be taken merely to refer to the achievement of full employment, but must also make reference to the achievement of low and stable inflation' (Meade 1978b). He argued that it is not sufficient to rely on incomes policy, of the conventional kind which was still fashionable in Britain. The fundamental problem is that a commitment to 'full employment' removes the threat of unemployment as a response to over-rapid wage increases, and it is on

this threat which wage and price stability in part depends. As a result, Meade argued that Keynesian policies should be 'stood on their head'. Demand management policy should be responsible for the maintenance of a slow and restrained rate of growth of money incomes, so as to put a 'lid' upon inflationary pressures. Incomes policy, or, more generally, the 'reform of wage-fixing', should be used – he argued – not to hold down prices but to promote employment.

This lecture contained three striking claims.

First, Meade's assertion that demand management would, inevitably, be excessively expansionary and would thereby promote inflation was essentially the same claim as that made the following year by Kydland and Prescott (1978) – a claim which went on to help them, too, to win the Nobel Prize. Meade's claim was made five years before the macroeconomic implications of the Kydland and Prescott idea were properly worked out by Barro and Gordon (1983).

Second, Meade's claim that macroeconomic policy should be confined to 'putting a lid on inflation' implied that employment would no longer be determined by a macroeconomic policy which was promoting full employment. As a result the levels of employment and of unemployment would be determined in some other way. At any point in time, said Meade, the 'reform of wage fixing' could be taken as given, and that would determine what we would now call the non-accelerating inflation rate of unemployment, or NAIRU. Meade then said that unemployment would gravitate towards the NAIRU, using the following argument. If unemployment was lower than the NAIRU, then inflation would be rising. But if the rate of growth of money incomes was effectively controlled by policy, then this would mean that policy would need to ensure that output fell, so as to prevent the growth of money incomes from rising above target. That would cause unemployment to rise towards the NAIRU, which would – in turn – stop inflation from rising. Meade used a similar argument to describe what would happen if unemployment was above the NAIRU. This line of reasoning effectively made Meade a follower of Friedman, who had claimed, in his fundamental paper published ten years earlier, that macroeconomic policy could not itself control the level of unemployment (Friedman 1968). Friedman's idea had been publicly broadcast in Great Britain, by Prime Minister Callaghan, in a famous speech given two years before Meade's lecture. But at the time this idea was too revolutionary for most macroeconomists in Britain. It was still widely thought that only monetarists believed something like this; Bob Rowthorn had caused uproar amongst the Cambridge Keynesians by claiming something of this kind just a year before Meade gave his lecture (Rowthorn 1977). Meade's lecture had the effect of detaching such a claim from its monetarist proponents, and began the process of making this claim mainstream in Britain, something which was eventually achieved by Layard et al. (1991).

Third, Meade discussed how, exactly, demand management policy (that is, fiscal and monetary policy) should be used to achieve the required slow and restrained growth of money incomes. His answer was that this should be done mainly by fine-tuned changes in tax rates, which, as mentioned above, he had discussed in *Consumers' Credit and Employment* (1938) and as he had suggested in his draft of the Full Employment White Paper in 1944. This answer made him very *unlike* Milton Friedman. In a subsequent mischievous talk to the Royal Economic Society, Meade (1981) created a taxonomy, so as to compare his new proposals with orthodox Keynesianism on the one hand and monetarism on the other. His mischief was to make the monetarists end up on the far left of his taxonomy, and to make the 'old-fashioned' Keynesians end up on the far right.

Meade presented a draft of this Nobel Prize lecture to the Marshall Society – Cambridge's student economics society. I was a research student in Cambridge at the time. As I recall, we did not know that Meade had just been awarded the Nobel Prize, or that what we were hearing was a dry run for his lecture in Stockholm. His lecture was a bit too un-Keynesian for me, and I stood up and said so. Meade defended his claims to the (large) audience, using the argument that policy works well when each policymaker is given an objective which he is likely to be able to

achieve – and that macroeconomic policymakers would be able to achieve a nominal income target, but would not be able to achieve an excessively optimistic employment target. (This was, again, a very Barro-Gordon-like answer.) But I had, by that time, read the papers by Phillips referred to above. So I stood up again and – rather bravely – said that, although this might be true, I thought that if his system was set up as a set of differential equations it would probably be unstable.

This question was to set in train a large research programme in the Department of Applied Economics in Cambridge. I had never met Meade before this lecture, but within a week he had asked me to work with him, and he then gradually gathered a large team to work with us, which included Andy Blake, Nicos Christodoulakis, Martin Weale and Peter Westaway, and also brought the control engineer Jan Maciejowski into the group. The resulting activity led to four substantial books (Meade 1982; Meade et al. 1983a; Meade 1986a; and Meade et al. 1989) and also to a number of tracts and articles in both technical and popular journals. Two main strands emerged in this work; we can describe these as being about inflation targeting and about supply side reforms.

Inflation Targeting

The second and fourth of the books just described set out Meade's proposed policy regime, in which there would be a target for nominal GDP, to be controlled primarily by means of changes in taxes. In Meade et al. (1983a) it was shown, using an estimated econometric model of the economy, that fine-tuned feedback rules for taxes really could be found which would keep nominal income close to a target path. The work used the multivariable control methods which Phillips (1957) had predicted would become available, which were supplied to the group by Jan Maciejowski.

This work culminated in Meade et al. (1989), called *Macroeconomic Policy: Inflation, Wealth and the Exchange Rate.* As a central part of the work for this book, Martin Weale oversaw the construction of an original empirical macroeconomic model, which developed the model being used at the National Institute of Economics Research in London at the time. It contained a number of rational-expectations features, which, at that time, were highly innovative. In particular, the model investment was driven by Tobin's q (that is, by the value of the stock market), which, following earlier work by Blanchard, was forward-looking, and which jumped in response to the expected future level of the interest rate (in exactly the same way as the exchange rate jumps in the Dornbusch model). And the model contained a forward-looking consumption function, with consumption partly depending on expected future income and thus on the expected future level of taxes. This is by now all rather familiar, but was ground-breaking at the time, although some of the underlying ideas had already been explained by Meade himself, nearly 20 years earlier, in *The Growing Economy* (Meade 1968).

Meade's policies were tried out on this model, using taxes as the policy instrument (and also the interest rate, for reasons explained below). This required the application of feedback control to a forward-looking model. That was necessary, given the rational-expectations features in the model, which made consumption and investment at any point in time depend on the expected level of taxes and the interest rate in the future, as has been explained above. The new ideas necessary for this work were developed jointly by the group in Cambridge, by a group in London led by David Currie and Paul Levine, and by Marcus Miller and Willem Buiter in Warwick and Bristol. The central idea driving that work on control methods was that rule-bound policies are necessary to guide an economy, if the world is forward-looking, since what economic agents do now depends on what they expect policy to do in the future. Such ideas were largely put on one side in the early to mid-1990s, when inflation-target regimes were first analysed theoretically, using simple backward-looking models (see, for example, Bean 1998.) But many of them have re-emerged in more recent technical work on targeting inflation in forward-looking, dynamic economies. For example, the idea of 'stabilization bias', which was understood very clearly by this group of people in England in the mid-1980s, was rediscovered and made popular by Michael Woodford nearly 15 years later, in the late 1990s.

Meade's young colleagues came to experience his skill at running a group of researchers – which I have begun to think he partly inherited from his experience in the Cambridge Circus so many years previously. As he passed 80 years of age, Meade presided over a weekly programme of meetings, at which his research group discussed the rational-expectations developments which I have described above. The day after each meeting, Meade would sit at home, in his village outside Cambridge, and write down an algebraic formulation of what we had all discussed. He would then walk to his local post office and send us a letter containing a photocopy of these handwritten notes. We would all then analyse his algebra and diagrams, in preparation for the next week's meeting.

It is fair to say that the policy proposals, which Meade's group developed, have not withstood the test of time. There are two explanations for this.

First, we now target the rate of inflation, not nominal incomes; Meade's nominal-income target regime was, effectively, only a precursor to the inflation-target regime which is now in place in the UK. Meade proposed a nominal income target, in part, because it was inherently more flexible than a *rigid* inflation target. It did not require that the inflation rate be exactly pinned down to an exactly pre-announced rate, but instead allowed, as explained above, for a (temporary) increase in inflation to be met by a (temporary) reduction in output, so as to ensure – on balance – that there would be no change in the rate of growth of nominal incomes. We now know that a *flexible* inflation-target regime is better than such a nominal-income target regime. But it took some years of research for us to understand just why this is so, work which is described, for example, in Hall and Mankiw (1993), Leiderman and Svensson (1995), and Woodford (2003). We have realized that the chief disadvantage of a nominal income target is that it does not 'let bygones be bygones': it requires that any overshoot which has occurred in the *level* of prices be clawed back, by means of a recession and lower subsequent inflation. But – at the same time – we have also realized that significant institutional development is required if one is to move from a purely rule-based system, like a nominal-income-target regime, to something like a rule-based but flexible inflation-target regime. To do this requires that the macroeconomic policymaking authorities be shielded from political influences which might force them to use their flexibility in an over-inflationary manner.

Second, we now use changes in interest rates, not changes in tax rates, in order to control inflation. We do this for three well-known reasons. First, it is easier to shield monetary policy from political influence than it is to do this for fiscal policy. Second the interest rate can be changed more regularly than taxes can be changed, and more quickly in response to new information – although fiscal procedures are less inflexible in some countries (such as the UK and New Zealand) than in others (such as the United States). Third, in an open economy monetary policy can have effects beyond those which can be caused by changes in taxes, because it can cause changes in a country's exchange rate which can, in turn, cause movements in exports and imports. This allows a country to externalize some of the costs of controlling shocks. That is a good idea if shocks in the world happen at different times in different countries.

Thus, as to both target and instrument, it appears that the world has moved on from Meade's proposals.

Nevertheless Meade's proposals had more general features, which *do* seem to have survived the test of time. Meade came to describe them as 'New Keynesian'. They were 'Keynesian' since, unlike Friedman, Meade continued to see the need for interventionist macroeconomic policies. (On this see Gordon 1990.) They were 'new' because Meade proposed a target for a nominal variable (nominal income) instead of having a target for real output. And, in addition, they were proposals for rule-bound policies. It is hard to remember how unusual, and how original, it was to combine these three features, in the early 1990s.

These three features of Meade's proposals seem to have had a significant influence on the development of macroeconomic policymaking in the UK in the early 1990s. It is also hard to remember just what a mess macroeconomic

policymaking was in Britain at that time. Following the country's brief flirtation with monetarism, it had joined the ill-fated ' exchange-rate mechanism' of the European Monetary System, which, in retrospect, appears to have been a pretty stupid policy framework. Following the UK's ejection from that system in September 1992, the Bank of England needed to quickly design a new policy regime. There was very little good theoretical guidance on what to do – other than Meade's. I say this, in particular, because the proposals of John Taylor, that monetary policy could follow a 'Taylor rule', really emerged only two years later (Taylor 1994). When the new regime was announced by the Bank, *within days,* it had Meade's three features – it was one in which interest rates would be actively used, in pursuit of a nominal variable (the inflation target), in a rule-bound (if flexible) manner. The outcome was one of the world's earliest inflation-targeting regimes (the British regime was second only to that established in New Zealand), a set-up which has developed into the world's best inflation-targeting system. I believe that *Macroeconomic Policy* (Meade et al. 1989) exerted some influence in the construction of this valuable new regime in Britain.

Furthermore, there are aspects of Meade's proposals which may yield further benefits in the future. *Macroeconomic Policy* suggests that policy should not just pursue a nominal anchor (taken to be a nominal income target in that book but it could just as well be an inflation target). The book also suggests that policy should pursue a target for the allocation of GDP between consumption and investment, so as to avoid 'selling off the family silver' (a phrase much discussed at the time), that is, so as to ensure that the supply side of the economy grows sufficiently rapidly. To do this, the book suggests that there should be rule-bound procedures for *two* policy instruments (both monetary policy – that is, interest-rate policy – and fiscal policy) in the joint pursuit of *two* targets (both the nominal anchor and the consumption-investment split). This was a very Meade-like suggestion in two ways: it synthesized a number of different ideas being discussed at the time, and it was characteristically complex and difficult to investigate (making this aspect of *Macroeconomic Policy* quite hard to understand).

One might argue that, in most circumstances, interest-rate policy can adequately control inflation, in the short to medium run, leaving fiscal policy to be more gradually adjusted so as to being about any desired changes in the consumption–investment mix, in the longer term. This is, for example, how the current British macroeconomic policymaking framework operates. In such cases monetary policy and fiscal policy can be considered separately, and a complex analysis of two instruments in pursuit of two targets is positively unhelpful.

Nevertheless, practical experience – in the United States, Japan, Europe and Australia – has shown that there are circumstances in which fiscal policy may need to assist in the pursuit of the inflation target, particularly when there are large falls, or increases, in demand. (See Garnaut 2005.) And recent theoretical work has shown that there may be more general advantages if fiscal and monetary policymakers can rely on each other to act in appropriate ways. (See Allsopp and Vines 2005.) The problems which might arise if the monetary and fiscal authorities cannot do this, and act independently of each other, were examined in Meade's very last published journal article (Meade and Weale 1995b). These problems have arisen very seriously in the Eurozone, where the European Central Bank and European governments do not cooperate, but not much, if at all, in the United Kingdom, for reasons examined theoretically in Kirsanova et al. (2005).

The Reform of Wage Fixing

The second part of Meade's project considered measures to promote employment through the reform of wage fixing. These were described in Meade (1982, 1984a, 1986a, 1986b). Looking back, one can credit Meade with having helped to create a sea change in the 1980s in British discussion of how wages ought to be fixed. Gone entirely are the ideas of rigid, centralized policies to hold down wages and prices by centralized administration. In their stead are proposals for policies which reinforce market mechanisms and which have their major impact

as employment-creating rather than price-controlling devices (Layard 1986). Meade's own suggestions included proposals for arbitration, a wage inflation tax and profit sharing.

On profit sharing and related topics Meade had already written a number of papers, starting in 1972 with 'The Theory of Labour Managed Firms and of Profit Sharing'; and his views on this subject also become influential in Britain. He was sympathetic to the ideas about workers' remuneration espoused in Weitzman's book *The Share Economy* (1984). But his criticisms of Weitzman were also important. Profit sharing *might* have beneficial effects for macroeconomic stability, through encouraging greater flexibility of workers' remuneration. But it *might also* do the opposite, if workers who concede profit sharing also come to exert an influence on the employment decisions of their firms, and use this influence to restrict employment opportunities and raise their own wages.

Meade went on working in this second area, long after the group of those working on demand management broke up after the publication of *Macroeconomic Policy* in the late 1980s. An important driving force in this work, and something which I have not discussed adequately in this article, was Meade's interest in the reform of the social security system. Such reform might make it possible to reconcile an efficient labour market – which might necessitate a pay-bargaining system that delivered low wages to some people – with a distribution of income which was equitable and just.

The year 1995 saw the production of Meade's last book *Full Employment Regained* (Meade 1995a), in which he attempted a synthesis of his ideas on demand management and on supply side reforms, arguing that full employment was possible providing that the appropriate reforms were undertaken. This, as Atkinson and Weale (2000) say, brought his career full circle. That career began, and ended, with Meade being concerned about the waste of resources and misery generated by high levels of unemployment. The Institute of Fiscal Studies hosted a seminar at which the ideas in his book were discussed. This was his last public appearance. And it was a gathering of many of the people whom he had influenced throughout his long career.

Influence

There can be no doubt that the *Theory of International Economic Policy* had an enormous influence upon our discipline. Corden and Atkinson (1979) and Johnson (1978) pay eloquent tribute to this. What I have said above suggests that there are many other ways in which he has exerted considerable influence. However, it is true that Meade is not as visible as some others of his generation.

This is probably due to his difficult manner of writing. This meant that his books were not as widely read as they might have been. Immediately one must exempt from this blanket statement Meade's popular articles and semi-popular tracts, which were beautifully written, and which displayed Meade's classical training to great effect, at the same time as being very persuasive. However his form of exposition, when he was doing fundamental economic theory, was very different. His 'style of work and presentation consists in the development of a *general* mathematical model of a problem, followed by translation of analysis of the various possible cases into literary English illustrated at most by arithmetical examples or simple diagrams' (Johnson 1978, p. 65; emphasis added). Johnson went on (p. 66) to complain about his 'taxonomic approach and dependence on rather inelegant personal mathematics'. This means, said Johnson, that 'students find it incredibly tedious to read his books and [find it] difficult to convince themselves that the effort is worthwhile in terms of the knowledge gained' (p. 65).

Corden and Atkinson made similar complaints, specifically about Meade's *Theory of International Economic Policy*, but by implication about his other work as well:

> . . . the . . . model of *The Balance of Payments* was very influential and . . . had a rapid impact on key writers and policy makers in the field. . .. By contrast, the influence of Trade and Welfare was more delayed, and to a great extent many of its original ideas were rediscovered later . . . Both books . . . are

> written in a taxonomic and rather heavy style, with no footnote references to the literature and a failure to highlight the author's original contributions. Although the books are immensely rewarding to serious students, their messages often reach a wider audience only through the intermediation of more succinct, if less original, writers. (Corden and Atkinson 1979, p. 530)

Elsewhere Corden and Atkinson talk of Meade's 'distinctive literary-arithmetical style', which now seems somewhat old fashioned compared with modern concise, simple algebraic expositions. Johnson (1978, p. 74) sums up the complaints, talking about the 'reader-repellent character of Meade's literary- arithmetical-cum-idiosyncratic-mathematical-appendix style of presentation'.

All this enables one to see why the spread of Meade's ideas had to rely, more than is usual, on his personal influence over his colleagues. It is easy to see how, in such circumstances, his influence could be underrated.

Yet one can easily see, too, why Meade deliberately chose to work in the way just described. It was his prodigious power to generalize – to see competing theories about any subject as part of a yet larger encompassing scheme of things – which caused him to create his vast architectural structures of taxonomy. These put off many readers. But, the structures having been created, a dedicated band of followers managed to climb up onto them, and then – when they came down again – to explain what they had seen to the rest of the profession. It was obviously easier to do this for those disciples and colleagues who had the good fortune to work directly with Meade, for they were able to discuss the insights of his work with him, as they worked through it. As will be clear, there were many such disciples and colleagues throughout Meade's long career. It is mainly through them, and thus mainly indirectly, that his influence spread so far.

Those who worked with Meade shared in his zest for life and in his acute sense of fun, some of which may be apparent to the reader of this account. They saw, too, his respect for careful argument, and his pleasure in a slow and measured conversation, through which such argument can be developed. But, especially, Meade conveyed to them his sense of the underlying moral purpose of our discipline. It is appropriate to end this assessment of Meade's work by discussing his views on that subject.

Underlying Philosophy

I mentioned in the introduction that Meade took up the study of economics because he wanted to help make the world a better place for ordinary men and women. In this he stood in the great Cambridge tradition of secular moralists, who might in the early Victorian age have become priests, but under the later influences of Darwinism and religious doubt turned instead to social improvement. The first volume of Skidelsky's biography of Keynes (Skidelsky 1983) links Keynes back to Marshall and Sidgwick in that enterprise. Meade, in turn, emphasized this shared objective as 'the decisive factor in binding me so closely to [Keynes] … he had … a passionate desire to devise a better domestic and international society' (Meade 1983b, p 268).

What is the conception of this better society that Meade strove for? And what is the role of the economist in helping to create it? The following few paragraphs, taken from a review article which he wrote in the late 1940s (Meade 1948b, p. 34), summarize some of his key ideas with a deceptive simplicity.

Meade writes that one's overall purpose is that 'of combining freedom, efficiency and equity in social affairs . . .'

> Two points should, however, be emphasised. First, this does not beg the question of planning. There may well be occasions . . . on which the State should rightly prepare general programmes for far-reaching structural changes in the use of the community's resources; and there may be sections of the economy (such as public investment) where the State should on all occasions plan ahead. But where planning takes place, it is still possible to use money and prices as a main, if not the main, instrument for getting the plan carried out.
>
> Secondly, there is no suggestion that on those occasions in which money and prices have been extensively used in the past the arrangements have been satisfactory. Far from it. In order that money

> and prices may fulfil their purpose three main conditions must be fulfilled. First, the total supply of monetary counters must be neither too great nor too small in relation to the total supply of goods and services to be purchased. Secondly, the total supply of monetary counters must be equitably distributed so that no one obtains more than a fair share of command over resources. Thirdly, no private person or body of persons must be allowed to remain in a sufficiently powerful position to rig the market for his own advantage.
>
> These conditions have not been fulfilled in the past. On the contrary, considerable state planning and much state intervention is required to ensure that these conditions are fulfilled. If, however, we wish to combine freedom, efficiency and equity in our economic life, we should proceed to make arrangements to see that these fundamental conditions are satisfied; and as they are more and more nearly fulfilled we should make a progressively greater use of the monetary and pricing systems. . ..

These 'fundamental conditions' have indeed been more nearly satisfied, in OECD countries, in the several decades since Meade wrote these words. But there is still much work to do. We still need to design intelligent monetary and pricing systems to deal with pressing global microeconomic problems (such as the threat of global warming, or the miserable health of the world poorest people), and with pressing global macroeconomic problems (such as large international imbalances, and the risks of financial crises in emerging market economies). It remains Meade's challenge to economists that we should develop policymaking institutions, and pricing systems, to deal with these problems, in ways which combine all of freedom, efficiency and equity, as much as possible.

See Also

- ▶ Absorption Approach to the Balance of Payments
- ▶ Elasticities Approach to the Balance of Payments
- ▶ Heckscher–Ohlin Trade Theory
- ▶ Inflation Targeting
- ▶ International Monetary Fund
- ▶ World Trade Organization

Selected Works

This article is a development of my entry about Meade in the first edition of *The New Palgrave: A Dictionary of Economics* (Vines 1987). That piece made use of two full-scale assessments of his work by Harry Johnson (1978) and by Corden and Atkinson (1979), and also used the short account by Harcourt (1985). In writing this piece I have been helped by a number of full-scale assessments of Meade's work which have appeared since 1987, by Atkinson (1996), Atkinson and Weale (2000), and Howson (2000). I have also referred to shorter accounts in Solow (1987), Greenaway (1990), and Corden (1996a, b). There is a bibliography of Meade's work attached to Johnson (1978), which was complete when it was assembled, except for 'a number of ephemeral newspaper articles and reviews'. From 1978 until his death in 1995, Meade continued to publish a remarkable amount, and a full bibliography is attached to Howson (2000). Accordingly, in the list below I include only those pieces of Meade's work to which I have referred explicitly, plus a few other particularly important pieces to which I do not refer. The bibliography, of course, includes all the pieces by other authors to which I make reference.

1933. *The rate of interest in a progressive state*. London: Macmillan.

1934. The amount of money and the banking system. *Economic Journal* 4: 98–107.

1936. *An introduction to economic analysis and policy.* London: Oxford University Press. 2nd ed., 1937. American edition, ed. C.J. Hitch with an Introduction by A.J. Hanson, New York: Oxford University Press, 1938.

1937. A simplified model of Mr Keynes' system. *Review of Economic Studies* 4: 98–107.

1938. *Consumers' credits and unemployment.* London: Oxford University Press.

1940. *The economic basis of a durable peace.* London: Oxford University Press.

1941. (With J.R.N. Stone.) The construction of tables of national income, expenditure, savings

and investment. *Economic Journal* 51: 216–233.
1944. (With J.R.N. Stone.) *National income and expenditure.* London: Oxford University Press.
1948a. *Planning and the price mechanism: The liberal socialist solution*. London: Allen & Unwin; New York: Macmillan, 1949.
1948b. Planning without prices. *Economica NS* 15: 28–35.
1951–5. *The theory of international economic policy,* vol. 1: *The balance of payments* (1951), vol. 2: *Trade and welfare* (1955a); With mathematical supplements. London/ New York: Oxford University Press.
1952. *A geometry of international trade.* New York: A. Kelley, 1969. London: Allen & Unwin.
1953a. *The Atlantic community and the dollar gap.* London: Friends of the Atlantic Union.
1953b. *Problems of economic union.* The Charles R. Walgreen Foundation Lectures. Chicago: University of Chicago Press.
1955b. *The theory of customs unions.* The De Vries Lectures, vol. 1. Amsterdam: North-Holland.
1955c. The case for variable exchange rates. *Three banks review* 27(September), 3–27. Repr. in *Readings in money, national income and stabilization policy,* ed. W. Smith and R. Teigen. Chicago: Richard D. Irwin, 1965.
1956a. *The Belgium–Luxembourg economic union, 1921–39: Lessons from an early experiment.* Essays in international finance no. 25. Princeton: International Finance Section.
1956b. *Japan and the general agreement on tariffs and trade.* The Joseph Fisher Lectures in Commerce, Adelaide University. Adelaide: Adelaide University Press.
1957. (With E.A. Russell.) Wage rates, the cost of living and the balance of payments. *Economic Record* 33(April): 23–28.
1958. *The control of inflation.* Inaugural lecture, Cambridge University. London: Cambridge University Press.
1961a. (With others.) *The economic and social structure of Mauritius.* Report to the Governor of Mauritius. London: Methuen.
1961b. *A neo-classical theory of economic growth.* London: Allen & Unwin. 2nd ed., 1964.
1964. *Efficiency, equality and the ownership of property.* London: AlIen & Unwin.
1965. *Principles of political economy: I. The stationary economy.* London: Allen & Unwin; Chicago: Aldine Press.
1965. (With F.H. Hahn.) The rate of profit in a growing economy. *Economic Journal* 75: 445–448.
1966. The outcome of the Pasinetti process: A note. *Economic Journal* 76: 161–165.
1968. *Principles of political economy: 1I. The growing economy.* London: Allen & Unwin; Chicago: University of Chicago Press.
1971a. *Principles of political economy: III. The controlled economy.* London: Allen & Unwin.
1971b. *Wages and prices in a mixed economy*, Occasional paper no. 35. London: Institute of Economic Affairs.
1972. The theory of labour managed firms and of profit sharing. *Economic Journal* 82: 402–428.
1973a. *The theory of economic externalities.* Geneva: Institut Universitaire des Hautes Etudes Internationales.
1973b. Economic policy and the threat of doom. In *Resources and population,* ed. B. Benjamin, P. Cox and J. Peel. London: Academic Press.
1973c. The inheritance of inequalities: Some biological, demographic, social and economic factors. *Proceedings of the British Academy* 59, Oxford: Oxford University Press.
1975a. *The intelligent radical's guide to economic policy.* London: Allen & Unwin.
1975b. The Keynesian revolution. In *Essays on John Maynard Keynes,* ed. M. Keynes. London: Macmillan.
1976. *Principles of political economy: IV. The just economy.* London: Allen & Unwin.
1977a. Autobiography. In *Nobel Lectures: Economics 1969–80,* ed. A. Lindbeck. Singapore: World Scientific Publishing Co, 1992. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1977/meade-autobio.html. Accessed 27 Mar 2007.

1977b. Banquet Speech on award of the Nobel Prize. In *Les Prix Nobel. The Nobel prizes 1977,* ed. W. Odelberg. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1977/meade-speech.html. Accessed 27 Mar 2007.
1978a. (With others.) *The structure and reform of direct taxation.* London: Allen & Unwin.
1978b. The meaning of internal balance. *Economic Journal* 88: 423–435.
1981. Comment on the papers by Professors Laidler and Tobin. *Economic Journal* 91: 49–55.
1982. *Stagflation,* vol. I: *Wage fixing.* London: Allen & Unwin.
1983a. (with D. Vines and J. Maciejowski.) *Stagflation,* vol. II: *Demand management.* London: Allen & Unwin.
1983b. Impressions of Maynard Keynes. In Worswick and Trevithic (1983).
1984b. A new Keynesian Bretton Woods. *Three Banks Review* 142 (June): 8–25.
1984c. A new Keynesian approach to full employment. *Lloyds Bank Review* 150: 1–18.
1984a. *Wage fixing revisited*, Occasional paper no. 72. London: Institute of Economic Affairs.
1986a. *Alternative systems of business organization and workers' remuneration.* London: Allen & Unwin.
1986b. *Different forms of share economy.* London: Public Policy Centre.
1988–90. *The collected papers of James Meade volumes I–IV,* (Volume IV jointly edited with D. Moggridge). London: Unwin Hyman.
1989. (With M. Weale, A. Blake, N. Christodoulakis and D. Vines.) *Macroeconomic policy: Inflation, wealth and the exchange rate*. London: Unwin and Hyman.
1993. The relation of Mr Meade's relation to Kahn's multiplier. *Economic Journal* 103: 464–465.
1995a. *Full employment regained*, Occasional paper no. 61. Department of Applied Economics, Cambridge University.
1995b. (With M. Weale.) Monetary union and the assignment problem. *Scandinavian Journal of Economics* 97: 201–222.

Bibliography

Alexander, S. 1952. The effects of a devaluation on a trade balance. *IMF Staff Papers* 2: 263–278.
Allsopp, C., and D. Vines. 2005. The macroeconomic role of fiscal policy. *Oxford Review of Economic Policy* 21: 485–508.
Atkinson, A. 1996. James Meade's vision: Full employment and social justice. *National Institute Economic Review* 157 (July): 90–97.
Atkinson, A., and M. Weale. 2000. James Meade: A memoir. In *1999 Lectures and Memoirs.* London: Published for the British Academy by Oxford University Press.
Barro, R., and D. Gordon. 1983. A positive theory of monetary policy in a natural rate model. *Journal of Political Economy* 91: 589–610.
Bean, C. 1998. The new UK monetary arrangements: A view from the literature. *Economic Journal* 108: 1795–1809.
Bhagwati, J., and V.K. Ramaswami. 1963. Domestic distortions, tariffs, and the theory of optimum subsidy. *Journal of Political Economy* 71: 44–50.
Brigden, J., D. Copeland, E. Dyason, L. Giblin, and C. Wickens. 1929. *The Australian tariff: An economic enquiry.* Melbourne: Melbourne University Press.
Corden, W.M. 1984. Booming sector and Dutch disease economics: Survey and consolidation. *Oxford Economic Papers* 36: 359–380.
Corden, W.M. 1996a. Special profile: James Meade, 1907–1995. *Review of International Economics* 4: 382–386.
Corden, W.M. 1996b. James Meade 1907–1995. *Economic Record* 72: 172–174.
Corden, W.M., and A. Atkinson. 1979. Meade James E. In *International enyclopaedia of the social sciences bibliographical supplement,* ed. D.L. Sills, vol. 18. New York: Free Press; London: Macmillan.
Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 57: 297–308.
Durbin, E. 1985. *New Jerusalems*. London: Routledge & Kegan Paul.
Fleming, J.M. 1951. On making the best of balance of payments restrictions on imports. *Economic Journal* 61: 48–71.
Fleming, J.M. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9: 369–379.
Friedman, M. 1953. The effects of a full employment policy on economic stability: A formal analysis. In *Essays in Positive Economics*. Chicago: Chicago University Press.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.
Garnaut, R. 2005. Is macroeconomics dead? Monetary and fiscal policy in historical context. *Oxford Review of Economic Policy* 21: 524–531.

Gordon, R. 1990. What is new Keynesian economics? *Journal of Economic Literature* 28: 1115–1171.
Greenaway, D. 1990. The intelligent radical on economic policy: An essay on the work of James Meade. *Scottish Journal of Political Economy* 37: 288–298.
Gregory, R. 1976. Some implications of the growth of the mineral sector. *Australian Journal of Agricultural Economics* 20 (2): 71–91.
Hall, R., and G. Mankiw. 1993. Nominal income targeting, NBER Working paper no 4439, Published in *Monetary Policy,* ed. G. Mankiw. Chicago: University of Chicago Press, 1994.
Harberger, A.C. 1950. Currency depreciation, income and the balance of trade. *Journal of Political Economy* 58: 47–60.
Harcourt, G.C. 1982. In *The social science imperialists: Selected Essays*, ed. P. Kerr. London: Routledge & Kegan Paul.
Harcourt, G.C. 1985. Meade, James (1907–). In *The social science encyclopaedia*, ed. A. Kuper and J. Kuper. London: Routledge & Kegan Paul.
Harrod, R.F. 1933. *International economics*. London: Nisbet.
Hicks, J. 1937. Mr Keynes and the classics: A suggested simplification. *Econometrica 5*: 147–159.
Howson, S. 2000. James Meade. *Economic Journal* 110: 122–145.
Johnson, H.G. 1951. The taxonomic approach to economic policy. *Economic Journal* 61: 812–832.
Johnson, H.G. 1965. Optimal trade intervention in the presence of domestic distortions. In *Trade growth and the balance of payments*, ed. R. Baldwin. Amsterdam: North-Holland.
Johnson, H.G. 1978. James Meade's contribution to economics. *Scandinavian Journal of Economics* 80: 64–85.
Kahn, R. 1931. The relation of home investment to unemployment. *Economic Journal* 41: 173–198.
Kahn, R. 1984. *The making of Keynes' general theory.* Cambridge: Cambridge University Press.
Keynes, J.M. 1930. *A treatise on money.* London: Macmillan.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.
Keynes, J.M. 1939. Relative movements of real wages and output. *Economic Journal* 49: 34–51.
Keynes, J.M. 1940. *How to pay for the war.* Repr. in Keynes (1971–88, vol. 9). London: Macmillan.
Keynes, J.M. 1971–88. *The collected writings of John Maynard Keynes*. London: Macmillan.
Kirsanova, T., J. Stehn, and D. Vines. 2005. The interactions between fiscal policy and monetary policy. *Oxford Review of Economic Policy* 21: 532–564.
Kydland, F.E., and E.C. Prescott. 1978. Rules rather than discretion. *Journal of Political Economy* 84: 473–491.
Laursen, S., and L. Metzler. 1950. Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32: 281–299.
Layard, R. 1986. *How to beat unemployment*. Oxford: Oxford University Press.
Layard, R., S. Nickell, and R. Jackman. 1991. *Unemployment: Macroeconomic performance and the labour market*. Oxford: Oxford University Press.
Leiderman, L., and L. Svensson, eds. 1995. *Inflation targets*. London: CEPR.
Minister of Reconstruction. 1944. *Employment policy.* Cmd. 6527. London: HMSO.
Mundell, R.A. 1962. The appropriate use of fiscal and monetary policy for internal and external stability. *IMF Staff Papers* 9: 70–77.
Phillips, C. 1920. *Bank Credit*. New York: Macmillan.
Phillips, A.W. 1950. Mechanical models in economic dynamics. *Economica* 17: 283–305.
Phillips, A.W. 1954. Stabilization policy in a closed economy. *Economic Journal* 64: 290–323.
Phillips, A.W. 1957. Stabilization policy and the time form of lagged responses. *Economic Journal* 67: 265–277.
Robinson, J. 1937. The foreign exchanges. In *Essays in the theory of employment.* London: Macmillan. 2nd ed., 1947.
Rowthorn, R. 1977. Conflict, inflation and money. *Cambridge Journal of Economics* 1: 215–239.
Samuelson, P. 1948. *Economics*. New York: McGraw Hill.
Shackle, G.L.S. 1967. *The years of high theory.* Cambridge: Cambridge University Press.
Skidelsky, R. 1983. *John Maynard Keynes: Hopes Betrayed 1883–1920*. London: Macmillan.
Skidelsky, R. 2000. *John Maynard Keynes: Fighting for Britain 1937–1946*. London: Macmillan.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Solow, R.M. 1987. James Meade at eighty. *Economic Journal* 97: 986–988.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Swan, T. 1963. Longer run problems of the balance of payments. In *The Australian economy,* ed. H.W. Arndt and W.M. Corden. Melbourne: Cheshire. Repr. in *Readings in international economics,* ed. R. Caves and H. Johnson. Homewood: Irwin, 1968.
Taylor, J. 1994. *Macroeconomic policy in a world economy: From econometric design to practical operation*. New York: Norton.
Van Dormael, A. 1978. *Bretton Woods: Birth of a monetary system*. London: Macmillan.
Viner, J. 1950. *The customs union issue*. New York: Carnegie Institute for International Peace.
Vines, D. 1987. Meade James Edward. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 3. London: Macmillan.
Vines, D. 1994. Unfinished business: Australian protectionism, Australian trade liberalisation and APEC. Shann Memorial Lecture, University of Western Australia. In *Australian macroeconomic policy debates: Contributions from the Shann memorial*

lectures, 1991–2000, ed. P.L. Crompton. Perth: University of Western Australia Press, 2005.
Vines, D. 1996. The Phillips machine. In *A.W. Phillips: Collected writings in contemporary perspective*, ed. R. Leeson. Cambridge: Cambridge University Press.
Vines, D. 2003. John Maynard Keynes, 1937–1946: The creation of international macroeconomics. *Economic Journal* 113: F338–F361.
Weitzman, M. 1984. *The share economy*. Cambridge, MA: Harvard University Press.
Williamson, J. 1983a. Keynes and the international economic order. In Worswick and Trevithick (1983).
Williamson, J. 1983b. *The exchange rate system*. Washington, DC: Institute for International Economics.
Wilson, T. 1982. Planning for the war and for the peace. In *Keynes as a policy adviser*, ed. A.P. Thirlwall. London: Macmillan.
Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.
Worswick, D., and D. Trevithick, eds. 1983. *Keynes and the modern world*. Cambridge: Cambridge University Press.
Young, W. 1987. *Interpreting Mr Keynes: The IS/LM Enigma*. Cambridge: Polity Press.

Mean Value

Chew Soo Hong

What is mean value? Conventional wisdom tells us that it represents, typifies or in some way measures the central tendency of a distribution. Familiar examples of mean value include the median, mode, arithmetic mean, geometric mean, harmonic mean and root-mean-square or more generally the *r*th root of the *r*th moment of a positive random variable.

A mean value arises when the following question is asked. In examining the 'effect' due to a given distribution of some quantity of interest, what value if *equally distributed* would result in the same overall effect? The *mean moon*, for example, is a fictitious moon which moves around the earth with a uniform speed and in the same time as the real moon. Typically, the effect of interest is measured by an index corresponding to the context of application. The most common measure-the sum of the quantities concerned – gives rise to the arithmetic mean.

Certainty Equivalence and Representative Income

The arithmetic mean was long considered a good rule-of-thumb for ordering risky prospects. In what has come to be known as the St Petersburg paradox, individuals are offered a lottery that pays 2^n dollars if in a sequence of independent coin tosses, the first head appears on the *n*th trial. While its expected payoff is infinite, few would find it to be worth much more than a few dollars. To account for this discrepancy, Bernoulli proposed in 1738 the expectation of a 'moral worth function as an alternative to the arithmetic mean. In particular, he adopted a logarithmic model of moral worth which yields a finite geometric mean as the *certainty equivalence* –the sure outcome which is as attractive as a given monetary lottery-of the St Petersburg lottery.

In the literature on income inequality, the concept of an *equally-distributed-equivalent* or *representative* income was proposed by Kolm (1969), Atkinson (1970) and Sen (1973). The representative income is defined to be the level of income which if distributed equally would result in the same overall level of social welfare as the existing income distribution. Given a representative income m as a mean value, there is a corresponding *relative* (*absolute*) *index of income inequality* $I_R(I_A)$ given by:

$$I_R(F) = 1 - \frac{m(F)}{\mu(F)}, \tag{1}$$

and

$$I_A(F) = \mu(F) - m(F), \tag{2}$$

where $\mu(F)$ refers to the arithmetic mean of the given income distribution F. Most measures of income inequality can be written in terms of expression (1) or (2).

Examples of Quasilinear Means

For a vector $x = (x_1, \ldots, x_N)$, a natural class of effect indices is given by the sum $\sum_{i=1}^{N} \upsilon(x_i)$ of a continuous and strictly monotone function υ of the x_i's. The resulting mean value m_υ called the *quasilinear mean* is given by:

$$\upsilon[m_\upsilon(x)] = \frac{1}{N}\sum_{i=1}^{N}\upsilon(x_i). \tag{3}$$

Examples of mean values which are special cases of (3) include the earlier mentioned arithmetic mean $(\upsilon \equiv x)$ geometric mean $(\upsilon \equiv \log x)$ harmonic mean $(\upsilon \equiv 1/x)$ root-mean-square $(\upsilon \equiv x^2)$ and the rth moment mean $(\upsilon \equiv x^y)$

For a probability distribution function F, the quasilinear mean $m_\upsilon(F)$ is given by:

$$m_\upsilon(F) = \upsilon^{-1}\left[\int \upsilon(x)\,\mathrm{d}F(x)\right]. \tag{4}$$

The first axiomatic characterization of (3) was proved in 1930 by Nagumo and Kolmogorov independently. De Finetti extended their result in the following year to (4) for simple (finite) probability distributions on a compact interval. Characteristic properties of the quasilinear mean will be discussed under the next heading.

As a model of certainty equivalence, the quasilinear mean corresponds to the *expected utility hypothesis* with v as the von Neumann–Morgenstern utility function. In this sense, Ramsey (1926) and von Neumann and Morgenstern (1947) provided other independent axiomatizations of the quasilinear mean.

It was suggested in a pioneering paper of Dalton (1920) that any measure of income inequality has an underlying social welfare function, which he further assumed to be additively separable (i.e. utilitarian) and symmetric. Dalton's approach was made precise in Atkinson (1970) which is equivalent to adopting the quasilinear mean as a model of the representative income. Atkinson considered specifically the one-parameter class of relative measures based on the rth moment mean m_r as the representative income.

Properties of the Quasilinear Mean

We represent the given distribution of the quantity of interest by a probability distribution function on an interval J of the real line $\mathbb{R}$. The *support* of a probability distribution function F denoted by supp(F) consists of each point x such that every open set around x has positive mass. The smallest closed interval containing supp(F) is denoted by conv supp(F). The degenerate probability distribution function whose support consists of a single point x is denoted by δ_xA vector $(x_1, \ldots, x_N)\varepsilon J^n$ can be represented as a simple probability distribution $\sum_{i=1}^{N}(1/N)\delta_{xi}$ with equal probability $1/N$ assigned to each outcome x_i. We denote by $x_\uparrow = \left(x_{[1]}, \ldots, x_{[N]}\right)$ the *increasing rearrangement* of $\mathrm{x} = (x_1, \ldots, x_N)$

A *mean value* is defined to be any functional on D_J satisfying the following fundamental property.

Property I (Intermediate Value Property):
$\forall F \in D_J,\ m(F) \in$ conv supp(F).

Property I tells us that the mean value of a probability distribution function lies between its lowest realizable outcome and its highest realizable outcome. Consequently, it implies the following:

Property SC (Consistency with Sure Outcomes):
$\forall x \in J,\ m(\delta_x) = x.$

We will present additional properties of the quasilinear mean below. One such property is consistency with the *first-degree stochastic dominance* partial order denoted by $\leq^1$

Property FSD (First-Degree Stochastic Dominance):
$\forall F,\ \mathrm{G} \in D_J,\ F \leq^1 G\, m(F) \in\ \leq m(G).$

Property I is implied by Property FSD which is often taken to be a universal property of mean values. While it is appealing in the certainty equivalence context, there are other mean values that do not necessarily satisfy this property.

The following is a characteristic property of the quasilinear mean.

Property Q (Quasilinearity or Substitution): ∀F, G, H $\in D_J$, and $\forall\beta \in (0,1)$, $m(F) = m(G) \Rightarrow$

$$m(\beta F + [1-\beta]H) = m(\beta G + [1-\beta]H).$$

A sequence of probability distributions $\{F_n\} \subset D_j$ is said to *converge in distribution* (or *weakly converge*) to a probability distribution $F \in D_j$ denoted by

$$F_n \xrightarrow{\mathscr{D}} F$$

if $F_n(x) \to F(x) \forall x$ such that F is continuous at x. It is known that

$$F_n \xrightarrow{\mathscr{D}} F \text{ if and only if } \int_J f\,\mathrm{d}F_n \to \int_J f\,\mathrm{d}F$$

for every bounded and continuous function f on J. The following is a definition of continuity often used in utility theory and in statistics.

Property CD (Continuity in Distribution):

$$\text{If } F_n \xrightarrow{\mathscr{D}} F, \qquad \text{then } m(F_n) \to m(F)$$

Note that the quasilinear mean is continuous in the above sense if and only if υ is bounded on J. This mles out the arithmetic mean among others as being continuous. The quasilinear mean, however, always satisfies the following weaker notions of continuity.

Property CC (Compact Continuity):

$$\text{If } \{F_n\}_{n=0}^{\infty} \subset D_J \xrightarrow{\mathscr{D}} F \in D_J$$

and $\mathrm{supp}(F_n) \subset K$ for some compact $K \subset J$ then

$$m(F) = \mathrm{Lim}_{n\to\infty} m(F_n).$$

The truncation of a probability distribution $F \in D_J$ by an interval $K \subset J$ is denoted by F_K.

Property E (Extension): $\forall F \in D_J$, $m(F) = \mathrm{Lim}_{n\to\infty} m(F_{K_n})$ for any increasing family of compact sets $\{K_n\}$ whose limit is J.

Property CC and Property E are implied by Property CD. Property CC essentially requires Property CD to hold on D_K for any compact interval $K \subset J$ Property E defines the mean value of a distribution F without compact support by the limit of the sequence of mean values of the truncated distributions F_{Kn} if the limit does not depend on the particular choice of the K_n. In this sense, the arithmetic mean of the Cauchy distribution which has unbounded support does not exist.

In the income inequality literature, a distribution is more unequal than another if the more equal distribution is obtained as a result of a sequence of transfers from a higher income individual to a lower income individual. This idea was captured by Hardy et al. (1934) definition of the majorization partial order and the corresponding definition of Schur-concavity. In utility theory, Rothschild and Stiglitz (1970) extended the above to the notion of *mean-preserving-increase* in risk on $D_{[A,B]}$. More generally, we say that a distribution G *dominates* another distribution F in the *second degree*, denoted by $G \geq {}^2F$, if $\forall x \in J$,

$$\int_{-\infty}^{\infty} [G(z) - F(z)]\mathrm{d}z \leq 0,$$

with equality as x approaches ∞ In this sense, the dominated distribution is more risky (less equal) than G. Consider:

Property SSD (Second-Degree Stochastic Dominance): $\forall F,\ G \in D_J,\ F \geq {}^2G$ implies that $m(F) \geq m(G)$.

It is known that m_υ satisfies the above if and only if υ is concave. In the sequel, we will examine other mean values that possess some of the properties discussed here. The mode specifically does not satisfy any of the properties except Property I.

Weighted Quasilinear Means

In addition to the continuous and strictly monotone υ function in m_υ, we introduce a nonvanishing *weighting* function w. For each x_j in a vector

$x \in J^N$, we assign a weight of $w(x_i)$. The resulting mean value $m_{\upsilon w}$ called the *weighted quasilinear mean* is given by:

$$\upsilon[m_{\upsilon w}(x)]\sum_{i=1}^{N} w(x_i) = \sum_{i=1}^{N} \upsilon(x_i) w(x_i). \quad (5)$$

For $F \in D_J$, $m_{\upsilon w}$ can be expressed as:

$$\int_J \{\upsilon(x) - \upsilon[m_{\upsilon w}(F)]\} w(x) \mathrm{d}F(x) = 0. \quad (6)$$

An example of $m_{\upsilon w}$ in mechanics is the following. For a simple pendulum of length L, the period vibration T is given by $(2\pi/g^{1/2})\mathrm{L}^{1/2}$, where g is the acceleration due to gravity. In general, the period of vibration of a pendulum with radial mass distribution F is given by the same formula except that L is now the ratio of the second moment (moment of inertia) to the first moment, i.e., $L = \int x^2 \mathrm{d}F(x) / \int xF(x)$. The length L – the length of an *equivalent* simple pendulum which yields the same period of vibration-is an example of the $m_{\upsilon w}$ mean with $\upsilon \equiv w \equiv x$.

In general, we may define the *ratio moment mean* $m_{s,t}$ by restricting $m_{\upsilon w}$ to those with $\upsilon \equiv x^s$ and $w \equiv x^t$ It is easy to show $\forall F \in D_J$ that $m_{s,t}$ increases in s and in t. In addition to the equivalent length L, the popularly used coefficient of variation provides another context in which $m_{1,\,1}$ arises:

$$\text{Coefficient of Variation} = \left[m_{1,1} - \mu\right]^{1/2}. \quad (7)$$

Note that expression (7) yields an absolute measure of income inequality with an inequality- preferring (risk-preferring) $m_{1,1}$ as the model of the representative income.

Other uses of the $m_{s,t}$ moment mean include:

$m_{1,2}$ = standard deviation · coefficient of skewness, and

$m_{2,2}$ = standard deviation · coefficient of kurtosis $+3)^{1/2}$

The following property is shared by $m_{\upsilon w}$ and m_{υ}.

Property B (Betweenness): $\forall F,\ G \in D_{\mathrm{J}}$, and $\beta \in [0, 1]$,

$m(F) \le m(G) \Rightarrow m(\beta F + [1 - \beta]G) \in [m(F), m(G)]$.

While betweenness appears to be a natural property for mean values, we will consider shortly mean values that do not satisfy it. The following is a characteristic property of $m_{\upsilon w}$.

Property SI (Substitution-Independence): Suppose $\exists F,\ G,\ H \in D_J$, and $\beta,\ \gamma \in (0, 1)$ such that

$$m(F) = m(G) \neq m(H)$$

and

$$m(\beta F + [1 - \beta]H) = m(\gamma G + [1 - \gamma]H),$$

then $\forall H' \in D_J$,

$$m(\beta F + [1 - \beta]H') = m(\gamma G + [1 - \gamma]H').$$

Clearly, if w is continuous on J, then $m_{\upsilon w}$ is compact continuous (Property CC). It can be shown that $m_{\upsilon w}$ will be continuous in distribution if w and $\upsilon \cdot w$ are both bounded on J. Chew (1983) proved that $m_{\upsilon w}$ is the only mean value satisfying Properties SC, B, SI, CC and E. To be consistent with Property FSD (Property SSD), we futher require υ and w to satisfy: $\forall s \in J$,

$$w(\cdot)[\upsilon(\cdot) - \upsilon(s)] \text{ is increasing (concave)} \quad (8)$$

Rank-Dependent Quasilinear Mean

Symmetry of a social welfare function was thought to be necessary for an inequality measure to be *impartial* in the sense of being invariant with respect to permutations of incomes among individuals within the population. We consider here mean values that are not inherently symmetric. Impartially is attained by restricting our evaluations to rank-ordered income vectors. The median provides an immediate example of a

rank-dependent mean value. The Gini arithmetic mean m_{Gini} (Sen 1973) derived from:

$$\text{Gini index} = 1 - \frac{m_{\text{Gini}}}{\mu} \tag{9}$$

provides another example. For an income vector $x_{\uparrow} \in J^N$, the Gini mean is given by:

$$m_{\text{Gini}}(x_{\uparrow}) = \sum_{i=1}^{N} \frac{[2(N-i)+1]}{N^2} x_{[i]}. \tag{10}$$

For a distribution $F \in D_J$, we have that:

$$m_{\text{Gini}}(F) \equiv \int_J z \; \mathrm{d}\left\{1 - [1 - F(z)]^2\right\}. \tag{11}$$

The above is a special case of the *rank-dependent quasilinear mean* m_{υ}^{g} For a probability distribution $F \in D_J$ we have:

$$m_{\upsilon}^{g}(F) = v^{-1}\left[\int_J \upsilon(x) \mathrm{d} g[F(x)]\right], \tag{12}$$

where $g : [0,1] \to [0,1]$ is nondecreasing and υ is continuous and strictly monotone. The quasilinear mean results when g is the identity map. In statistics, the rank-dependent mean corresponds to the class of L-estimators. We tabulate in Table 1 a number of well known m_{v}^{g} means with $v \equiv x$ which we denote by, m^{g}

It is clear that m_{υ}^{g} satisfies Properties I, FSD, CC and E. If υ is bounded, then it will be continuous in distribution. In addition, it satisfies Property SSD if and only if υ and g are both concave (Chew et al. 1987).

We state a characteristic property of m_{υ}^{g} in the following. We say that x and y are *rank preserving* if, for each i,

$$x_{[i]} \in \left[y_{[i-1]}, y_{[i+1]}\right] \text{ and } y_{[i]} \in \left[x_{[i-1]}, x_{[i+1]}\right].$$

The pair (w, z) is said to be a *rank-preserving rearrangement* of (x, y) if w and z are rank- preserving and, for each i,

$$\left\{w_{[i]}, \; z_{[i]}\right\} = \left\{x_{[i]}, \; y_{[i]}\right\}.$$

Property CI (Commutative Independence): $\exists \alpha \in (0,1)$ such that $\forall p \in \Delta^{N-1}$, $x, y \in J^N$ with x and y being rank-preserving and $x_{\uparrow} \leq y_{\uparrow}$,

$$m\left\{\sum_{i=1}^{N} p_i \delta_{m[\alpha\delta_{x_{[i]}} + (1-\alpha)\delta_{y_{[i]}}]}\right\} = m\left\{\alpha\delta_m\left(\alpha\sum_{i=1}^{N} p_i \delta_{w_{[i]}}\right) + (1-\alpha)\delta_m\left(\sum_{i=1}^{N} p_i \delta_{z_{[i]}}\right)\right\} \tag{13}$$

for any rank-preserving arrangement (w, z) of (x, y).

Recently, Quiggin (1982) provided a generalization of expected utility for simple probability distributions which corresponds to the m_{υ}^{g} model with $g(1/2) = 1/2$. Yaari (1987) independently axiomatized a theory of preference corresponding to m_{υ}^{g} with $\upsilon \equiv x$ Chew (1985b) axiomatized the m_{υ}^{g} mean in terms Properties SC, FSD, CI, CC and E.

Mean Value, Table 1

m^g	Probability transformation function g
Median	Step function at $p = 1/2$
α - Winsorized mean	$g(p) = \begin{cases} 0 & p \in [0, \alpha] \\ p & p \in (\alpha, 1-\alpha) \\ 1 & p \in [1-\alpha, 1] \end{cases}$
α - trimmed mean	$g(p) = \begin{cases} 0 & p \in [0, \alpha] \\ p/(1-(1-2\alpha) & p \in (\alpha, 1-\alpha) \\ 1 & p \in [1-\alpha, 1] \end{cases}$
Gini mean	$g(p) = 1 - (1-p)^2$
s - Gini mean (Donaldson and Weymark 1980)	$g(p) = -(1-p)^s$

Implicit-Weighted Quasilinear Mean

The mean values introduced thus far are defined explicitly in terms of operations relative to the given probability distribution. We present here a general class of implicitly defined mean values which are closely related to the M-estimator in robust statistics proposed by Huber (1964). The *implicit-weighted quasilinear mean* $m_{\upsilon(\cdot)}w(\cdot,\cdot)$ is defined to be the solution of:

$$\int_J [\upsilon(x) - \upsilon(s)]w(x,s)\,\mathrm{d}F(x) = 0, \tag{14}$$

where $w(x,\ s)$ is nonvanishing and $[\upsilon(\cdot) - \upsilon(s)]w(\cdot,\ s)$ is strictly monotone for each s. We have the weighted quasilinear mean when $w(x,\ s) \equiv w(x)$ and the quasilinear mean when $w \equiv$ constant.

Huber (1964) proposed a class robust location estimators as the solution of

$$\int \phi(x-s)\,\mathrm{d}F(x) = 0. \tag{15}$$

The Huber estimators are special cases of $M_{\upsilon w}$ with $\upsilon \equiv x$ and $w(x,\ s)[x-s] \equiv \phi(x-s)$. Fishburn (1986) axiomatized the case of (14) with w symmetric, i.e., $w(x,s) = w(s,\ x)$ and for each $s \in J,\ w(x,s)[\upsilon(x) - \upsilon(s)]$ strictly monotone in x.

An alternative way to write (14) is given by the following. First we define the *weighted transformed probability distribution* $F^{w(\cdot,s)}$ by:

$$\mathrm{d}F^{w(\cdot,s)}(x) = w(\cdot,\ s)\mathrm{d}F(x)/\int_J w(y,s)\,\mathrm{d}F(y). \tag{16}$$

Then implicit-weighted quasilinear mean is the solution of

$$s = v^{-1}\left[\int_J \upsilon(x)\,\mathrm{d}F^{w(\cdot,s)}(x)\right]. \tag{17}$$

In the case of the weighted quasilinear mean where w does not depend on s, (17) has the simpler form:

$$m_{\upsilon w}(F) = \upsilon^{-1}\left(\int_J \upsilon\,\mathrm{d}F^w\right). \tag{18}$$

It can be shown based on Chew (1985a) that the implicit $m_{\upsilon w}$ mean is the most general class of mean values having the betweenness property in addition to Properties SC, CC and E. To satisfy FSD (SSD), we further require $w(\cdot,\ s)[\upsilon(\cdot) - \upsilon(s)]$ to be increasing (concave) for each $s. \in J$.

A General Form of Mean Value

There is a pattern in the preceding exposition. The mean value in each case is defined to be the solution of an equation of the following form:

$$\upsilon(s) = \int_J \upsilon \mathrm{d}\phi_s(F) \tag{19}$$

where for each $s \in J,\ \phi_s : D_J \to D_J$ is *support-attenuating*; $\forall F \in D_J$, conv supp$[\phi_s(F)] \subset$ conv supp(F). If $\forall F \in D_J,\ \mathrm{supp}(\phi_s(F) = \mathrm{supp}(F)$ then we say that ϕ_s is *support-preserving*. Most of the mean values are defined relative to a support-preserving ϕ_s. The exceptions include the median, the α-trimmed mean and the α-Winsorized mean. The 'weighted' mean values would lose their support-preserving property when the weight function is allowed to vanish within the interior of J.

A mean value is an *explicit* one if ϕ_s does not depend on s. Otherwise, we have an *implicit* mean value. It is clear that any functional defined by (19) satisfies the intermediate value property. Conversely, any mean value functional $m(F)$ can be written in terms of (19) via the degenerate support-attenuating map $F \to \delta_{m(F)}$. If we require compact continuity, then ϕ_s needs to be *continuous* in the sense that $\phi_s(F_n)$ converges in distribution whenever F_n does.

Table 2 tabulates the known mean values and some new ones. All of these mean values satisfy Properties SC, I, CC and E (with w continuous and nonvanishing on J and g continuous and strictly increasing). These together with the stated characteristic properties yield the corresponding axiomatic characterizations of the respective mean values. Note that the unaxiomatized mean values in Table 2 are obtained by performing

Mean Value, Table 2

Mean value	Probability transformation	Characteristic properties
Quasilinear mean	identify	Q, B or FSD
Weighted quasilinear mean	$F \to F^w$	SI, B
Rank-dependent quasilinear mean	$F \to g[F(\cdot)]$	CI, FSD
Rank-dependent weighted quasilinear mean	$F \to g[F^w(\cdot)]$	
Weighted rank-dependent quasilinear mean	$F \to \{g[F(\cdot)]\}^w$	
Implicit-weighted quasilinear mean	$F \to F^{v(\cdot,s)}$	B
Rank-dependent implicit weighted quasilinear mean	$F \to g\left[F^{w(\cdot,s)}(\cdot)\right]$	
Implicit-weighted-rank-dependent quasilinear mean	$F \to \{g[F(\cdot),s]\}^{w(\cdot,s)}$	
Mean value	$F \to \phi_s(F)$	

sequentially a 'weighting transformation' and a 'rank-dependent transformation' on the given distribution F. This is illustrated below:

$$F \to F^{w(\cdot,s)} \to g\left[F^{w(\cdot,s)}(\cdot)\right],$$

versus

$$F \to g[F(\cdot)] \to \{g[F(\cdot)]\}^{w(\cdot,s)}.$$

In general, we can compose a support-preserving maps by performing a sequence of such weighting and rank-dependent transformations with a number of w and g functions.

See Also

- ▶ Allais Paradox
- ▶ Expected Utility and Mathematical Expectation
- ▶ Expected Utility Hypothesis
- ▶ Utility Theory and Decision Theory

Bibliography

Atkinson, A.B.. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–263.

Bemoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialas Petropolitanae* 5: 175–192. Trans. 'Exposition of a new theory on the measurement of risk', *Econometrica* 22 (1954): 23–26.

Chew, S.H. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* 51: 1065–1092.

Chew, S.H. 1985a. Implicit weighted and semi-weighted utility theories, M-estimators, and non-demand revelation of second price auctions for uncertain auctioned objects, Working paper No. 155. Department of Political Economy, Johns Hopkins University.

Chew, S.H. 1985b. An axiomatization of the rank-dependent quasilinear mean generalizing the Gini mean and the quasilinear mean, Working paper No. 156. Department of Political Economy, Johns Hopkins University; revised 1986.

Chew, S.H., Kami, E. and Safra, Z. 1986. Risk aversion in the theory of expected utility with rank- dependent probabilities. Forthcoming in the *Journal of Economic Theory.*

Dalton, H. 1920. The measurement of inequality of incomes. *Economic Journal* 20: 348–361.

Donaldson, D., and J.A. Weymark. 1980. A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory* 22: 67–86.

de Finetti, B. 1931. Sul concetto di media. *Giornale dell' Intituto Italiano degliAttuari* 2: 369–396.

Fishburn, P.C. 1986. Implicit mean value and certainty equivalence. *Econometrica* 54: 1197–1205.

Hardy, G.H., J.E. Littlewood, and G. Polya. 1934. *Inequalities*. Cambridge: Cambridge University Press.

Huber, P.J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35: 73–101.

Kolm, S-Ch. 1969. The optimal production of social justice. In *Public economics*, ed. J. Margolis and H. Guitton. London/New York: Macmillan.

Kolmogorov, A. 1930. Sur la notion de la moyenne. *Rendiconti Accademia dei Lincei* 6(12): 388–391.

Nagumo, M. 1930. Uber eine Klasse der Mittelwerte. *Japan Journal of Mathematics* 7: 71–79.

von Neumann, J., and O. Morgenstem. 1947. *Theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press.

Quiggin, J. 1982. Anticipated utility theory. *Journal of Economic Behavior and Organization* 3: 323–343.

Ramsey, F.P. 1926. Truth and probability. In *The foundations of mathematics*, ed. R.B. Braithwaite. London: Routledge & Kegan Paul.

Rothschild, M., and J.E. Stiglitz. 1970. Increasing risk: I A definition. *Journal of Economic Theory* 2: 225–243.
Sen, A. 1973. *On economic inequality.* Oxford: Oxford University Press.
Yaari, M.E. 1987. The dual theory of choice under risk: Risk aversion without diminishing marginal utility. *Econometrica* 55: 95–115.

Meaningfulness and Invariance

Louis Narens and R. Duncan Luce

Abstract

Given a qualitative scientific structure, a structure preserving mapping into a numerical, vectorial, or geometric structure is called a *representation* of it. Within such numerical, vectorial or geometric structures other concepts can always be defined. Some of these correspond to a qualitative property of the underlying system, and they are called 'meaningful' concepts. And others do not correspond to a qualitative property; and they are called 'meaningless'. The article investigates precise meanings of 'meaningfulness' and 'meaninglessness' and their relation to several notions of invariance, some of which are widely used in science.

Keywords

Dimensional analysis; Interpersonal comparison of utilities; Invariance; Meaningfulness; Measurement; Representations; Scientific definability

JEL Classifications

B4

Few disavow the principle that scientific propositions should be meaningful in the sense of asserting something that is verifiable or falsifiable about the qualitative or empirical situation under discussion. What makes this principle tricky to apply in practice is that much of what is said is formulated not as simple assertions about qualitative or empirical events – such as a certain object sinks when placed in water – but as laws formulated in rather abstract, often mathematical, terms. It is not always apparent exactly what class of qualitative observations corresponds to such (often numerical) laws. Theories of meaningfulness are methods for investigating such matters, and invariance concepts are their primary tools.

The problem of meaningfulness, which has been around since the inception of mathematical science in ancient times, has proved to be difficult and subtle; even today it has not been fully resolved. This article surveys some of the current ideas about it, and illustrates, through examples, some of its uses. The presentation requires some elementary technical concepts of measurement theory (such as representation, scale type, and so on), which are explained in measurement, theory of.

Concepts of Meaningfulness

Some Notation and Definitions

The operation of functional composition is denoted *. The Cartesian product of $T_1, \ldots, T_n$ is denoted $\prod_i^n Ti$.

A *scale* $\mathscr{S}$ is a set of functions from a qualitative domain, a set X endowed with one or more relations, into the real numbers. Elements of $\mathscr{S}$ are called *representations*. An example is the usual physical scale to measure length. Two of its representations are the *foot representation* and the *centimeter representation.* $\mathscr{S}$ is said to be

- a *ratio scale* if and only if for each φ in $\mathscr{S}$,

$$\mathscr{S} = \{r\varphi | \, r > 0\},$$

- an *interval scale* if and only if for each φ in $\mathscr{S}$,

$$\mathscr{S} = \{r\varphi + s | \, r > 0, s \text{ a real}\},$$

- an *ordinal scale* if and only if for each φ in $\mathscr{S}$, the range of φ is a (possibly infinite) interval of reals and

$$\mathscr{S} = \{f * \varphi \backslash f \text{ is a strictly monotonic function from the range of } \varphi \text{ onto itself}\}.$$

Intuitive Formulation of Meaningfulness and Some Examples

The following example, taken from Suppes and Zinnes (1963), nicely illustrates part of the problem in a very elementary way. Which of the following four sentences are meaningful?

(i) Stendhal weighed 150 on 2 September 1839.
(ii) The ratio of Stendhal's weight to Jane Austen's on 3 July 1814 was 1.42.
(iii) The ratio of the maximum temperature today to the maximum temperature yesterday is 1.10.
(iv) The ratio of the difference between today's and yesterday's maximum temperature to the difference between today's and tomorrow's maximum temperature will be 0.95.

Suppose that weight is measured in terms of the ratio scale $\mathscr{W}$ (which includes among its representations the pound and kilogram representations and all those obtained by just a change of unit), and that temperature is measured by the interval scale $\mathscr{T}$, which for this example includes the Fahrenheit and Celsius representations. (The Kelvin representation for temperature, which assumes an absolute zero temperature, is not in $\mathscr{T}$.) Then Statement (*ii*) is meaningful, because with respect to each representation in $\mathscr{W}$ it says the same thing, that is, its truth value is the same no matter which representation in $\mathscr{W}$ is used to measure weight. That is not true for Statement (*i*), because (*i*) is true for exactly one representation in $\mathscr{W}$ and false for all of the rest. Thus we say that (*i*) is 'meaningless'. Similarly, *(iv)* is meaningful with respect to $\mathscr{T}$ but (*iii)* is not.

The somewhat intuitive concept of meaningfulness suggested by these examples is usually stated as follows. Suppose a qualitative or empirical attribute is measured by a representation from a scale of representations $\mathscr{S}$. Then a numerical statement involving values of the representation is said to be *quantitatively meaningful* if and only if its truth (or falsity) is constant no matter which representation in $\mathscr{S}$ is used to assign numbers to the attribute. There are obvious formal difficulties with this definition, for example the concept of 'numerical statement' is not a precise one. More seriously, it is unclear under what conditions this is the 'right' definition of meaningfulness, for it does not always lead to correct results in some well- understood and non-controversial situations. (See the discussion involving situations where the measurement scale consists of a single representation for an example.) Nevertheless, it is the concept most frequently employed in the literature, and invoking it often provides insight into the correct way of handling a quantitative situation – as the following still elementary but somewhat less obvious example shows.

Consider a situation where M persons rate N objects (for example, M judges judging N contestants in a sporting event). For simplicity, assume that person i rates objects according to the ratio scale of representations $\mathscr{R}_i$. The problem is to find an ordering on the N objects that aggregates the judgements of the judges in a reasonable way. It can be shown that their judgements cannot be coordinated in such a way that, for all R_i in $\mathscr{R}_i$ and Rj in $\mathscr{R}_j$ that for some object a, the assertion $R_i(a) = R_j(a)$ is justified philosophically. The difficulties underlying such a coordination are essentially those that arise in attempting to compare individual utility functions. The latter problem – the 'interpersonal comparison of utilities' – has been much discussed in the literature, as for example in Narens and Luce (1983). It is generally agreed that there are great, if not insurmountable, difficulties in carrying out such comparisons. Any rule that does not involve coordination among the raters can be formulated as follows. First, let F be a function that assigns to an object the value $F(r_1,\ldots, r_M)$ whenever person i assigns the number r_i to the object. Second, assume that object a is ranked just as high as b if and only if the value assigned by F to a is at least as great as that assigned by F to b. In practice F is often taken to be the arithmetic mean of the ratings $r_1,\ldots, r_M$ (for example, Pickering et al. 1973). Observe, however, that arithmetic means for this kind of rating situation, in general, produce a non-quantatively meaningful ranking of objects, as illustrated by the following special case. Suppose $M = 2$ and, for $i = 1,2$, R_i is person's i

representation that is being used for generating ratings, and

$$R_1(a) = 2, \quad R_1(b) = 3, \quad R_2(a) = 3, \text{ and } R_2(b) = 1.$$

Then the arithmetical mean of the ratings for *a,* 2.5, is greater than that for *b*, 2, and thus *a* is ranked above *b*. However, meaningfulness requires the same order if any other representations of persons 1 and 2 rating scales are used, for example, $10R_1$ and $2R_2$. But for this choice of representations, the arithmetic mean of *a*, 13, is less than that of *b*, 16, and thus *b* is ranked higher than *a*.

It is easy to check that the geometrical mean of rankings for an object,

$$F(r_1, \ldots, r_M) = [r_1, \ldots, r_M]^{\frac{1}{M}},$$

gives rise to a quantitatively meaningful, non-coordinated rule for ranking objects. It can be shown under plausible conditions that all other meaningful, non-coordinated rules give rise to the same ranking as that given by the geometric mean (Aczél and Roberts 1989).

Many other applications of quantitative meaningfulness have been given by various researchers. In particular, Roberts (1985) provides a wide range of social science examples. In some contexts, quantitative meaningfulness presents certain technical difficulties that require some modification in its definition (see, for example, Roberts and Franke 1976; Falmagne and Narens 1983).

Meaningfulness and Statistics

Another area of importance to social scientists in which invariance notions are thought to be relevant is applying statistics to numerical data. The role of measurement considerations in statistics and of invariance under admissible scale transformations was first emphasized by Stevens (1946, 1951); this view quickly became popularized in numerous textbooks, and it produced extensive debates in the literature. Continued disagreement exists, mainly created by confusion arising from the following two simple facts:

- Measurement scales are characterized by groups of admissible transformations of the real numbers.
- Statistical distributions exhibit certain invariances under appropriate transformation groups, often the same groups (especially the affine transformations), as those that arise from measurement considerations.

Because of these facts, some scientists have concluded that the suitability of a statistical test is determined, in part, by whether or not the measurement and distribution groups are the same. Thus, it is said that one may be able to apply a test, such as a t-test, that rests on the Gaussian distribution to ratio or interval scale data, but surely not to ordinal data, because the Gaussian distribution is invariant under the group of positive affine transformations, $x \to rx + s$, r, s *real*, $r > 0$ – which arises in both the ratio and the interval case but not in the ordinal one. Neither half of the assertion is correct. First, a significance test should be applied only when its distributional assumptions are met, and they may very well hold for some particular representation of ordinal data. And second, a specific distributional assumption may well not be met by data arising from a particular scale of measurement. For example, reaction times, being times, are measured on a physical ratio scale, but they are rarely well approximated by a Gaussian distribution.

What is true, however, is that any proposition (hypothesis) that one plans to put to statistical test or to use in estimation had better, itself, be quantitatively meaningful with respect to the scale used for the measurements. In general, it is not quantitatively meaningful to assert that two means are equal when the quantities are measured by an ordinal scale, because equality of means is not invariant under strictly increasing transformations. Thus, no matter what distribution holds and no matter what test is performed, the result may not be quantitatively meaningful, because the hypothesis is not. In particular, if an hypothesis is about the measurement structure itself, for example that the representation is additive over a concatenation operation, then it is essential that the following propositions (*a*) and *(b)* hold, where a

symmetry of a structure is by definition an isomorphism of the structure onto itself: (a) the hypothesis be invariant under the symmetries of the structure and therefore invariant under the scale used to measure the structure. (Because it is assumed that scales of measurement are structure preserving functions from a qualitative structure onto a quantitative one, (*a*) immediately follows). And (*b*) the hypotheses of the statistical test be met without going outside the transformations of the measurement representation. See Luce et al. (1990) for a more detailed discussion of this issue.

Concepts of Invariance

Measurement laws are quantitative laws based primarily on interrelationships of scales of measurement. They have in common with quantitative meaningfulness that they are derived through considerations of admissible transformations of the measurements of relevant variables. In the view of Falmagne and Narens (1983, p. 298) they arise in an empirical situation 'that is governed by an empirical law of which we know little of its mathematical form and a little of its invariance properties, but a lot about the structure of the admissible transformations of its variables, and use this information to greatly delimit the possible equations that express the law'. They are generalizations of the kind of laws that have a longstanding tradition in physics, where they are known as laws derived according principles of 'dimensional analysis'. These principles involve the assertion that laws of nature are in a deep sense invariant under changes of unit, which correspond to invariance under symmetries. Thus, knowledge of the scale type of the relevant variables – a strong presupposition – greatly limits the forms of laws.

Measurement Laws: Simplest Case

These principles were introduced into the behavioural sciences by Luce (1959), which was concerned with special cases of 'possible psychophysical laws'. He generalized dimensional analysis, which only assumed ratio scale transformations of the several variables, to the more general situation of the measurement scale types described by S.S. Stevens. Luce (1964) extended the 1959 formulation to include a few important cases of a single function of many variables.

Luce (1959) considered the case where the independent variable x and the dependent variable y were related by a law, $y = f(x)$, where f was some continuous function. He assumed that this law was invariant under admissible transformations of measurements, that is, for each admissible transformation φ of the independent variable, there was an admissible transformation ψ of the dependent variable such that for all x and y,

$$y = f(x) \text{ iff } \psi(y) = f(\varphi(x)). \tag{1}$$

The following is an example of a use of Luce's theory. Suppose x is an objective variable measured by a ratio scale, for example, a physical variable such as the intensity of light or the weight of gold, and y is the subjective evaluation of x, for example, the subjective brightness of light, the subjective value of gold, and f is the law linking x and y. Suppose x and y are both measured on ratio scales and f is continuous. Suppose further that f satisfies Eq. 1. Under these conditions, Luce shows that there are real numbers r and a, a depending on ψ, such that

$$f(x) = ax^r. \tag{2}$$

His method of proof was to show that Eq. 1 implied that f satisfied the functional equation $h(s)f(t) = f(st)$ for some continuous function h and all positive s and t, and that this functional equation had Eq. 2 as its only solution.

For most applications, such as the above brightness and subjective value examples, the scale for the independent variable is known and continuity is a reasonable idealized approximation. Sometimes theory will specify the measurement scale for the dependent variable. However, often the scale for the dependent variable is unknown, and in many cases, unobservable, as, for example, when it is subjective. In such situations, the measurement scale for the dependent variable has to be hypothesized or derived from theory. It can be hypothesized to be one of several theoretically reasonable types of measurement

scales, and then methods similar to the one used to derive Eq. 2 can be used to arrive at a measurement law for each type of hypothesized scale. The set of resultant measurement laws provides a clear set of quantitative hypotheses for empirical testing. Quite often such hypotheses turn out to be a good place to begin a scientific investigation.

Measurement Laws: More Complex Cases

In a number of ways, Falmagne and Narens (1983) greatly generalized Luce's 1959 approach for deriving laws from measurement considerations. In particular:

- Instead of one independent variable and one dependent variable, they assumed n independent variables and one dependent variable. (They formulated matters for two independent variables to simplify notation, but their approach easily extends to n independent variables.)
- They allowed for a general relationship R among the admissible transformations of the independent variables to hold; that is, for the sets T_i of admissible transformations of the independent variables $x_1,\ldots,x_n$, R can be any nonempty subset of $\prod_i^n T_i$.
- They allowed for more general kinds of laws by allowing for a family $\mathscr{F}$ of functions to relate the dependent variable with n independent variables. They interpret $\mathscr{F}$ as follows. Initially, representations $\phi_1, \ldots, \phi_n$ are used to measure the n independent variables, $x_1,\ldots,x_n$. These measurements determine a function $f(\phi_1(x_1), \ldots, \phi_n(x_n))$ that is the value of the dependent variable measured on an unknown scale when $x_1,\ldots,x_n$ are measured by $\phi_1, \ldots, \phi_n$. There are other equally valid ways of measuring *each* independent variable x_i. These are obtained by transforming ϕ_i by the elements of T_i. However, valid measurements for the *set* of independent variables may be additionally constrained by the empirical law relating the dependent variable to the independent variables. The additional constraint is captured by the relation R. Thus each other valid measurement of the independent variables is given by $\tau_1 * \phi_1, \ldots, \tau_n * \phi_n$ for some $\tau_1, \ldots, \tau_n$ such that $R(\tau_1, \ldots, \tau_n)$. The law giving the numerical value of the dependent variable, when the set of independent variables $x_1,\ldots, x_n$ are measured respectively by $\tau_1 * \phi_1, \ldots, \tau_n * \phi_n$, is given by

$$f_{\tau_1,\ldots,\tau_n}(\tau_1 * \phi_1(x_1), \ldots, \tau_n * \phi_n(x_n)).$$

In this way, it is the family of functions,

$$\mathscr{F} = \{f_{\tau_1,\ldots,\tau_n}(\tau_1 * \phi_1(x_1), \ldots, \tau_n * \phi_n(x_n)) \mid R(\tau_1, \ldots, \tau_n)\},$$

that expresses the empirical law relating the dependent variable to the independent variables $x_1,\ldots, x_n$. Only in very restrictive cases will $\mathscr{F}$ consist of a single function.

Order Meaningfulness

In place of assuming the scale type of the dependent variable, they assume 'order meaningfulness', that is, they assume the following. Using the just presented notation, suppose $\mathscr{F}$ is a family of functions that is a law relating the dependent variable with n independent variables and $f_{\sigma 1, \ldots, \sigma n}$ and $f_{\tau_1,\ldots,\tau_n}$ are in $\mathscr{F}$. Then for all $x_1, \ldots, x_n$ and $u_1, \ldots, u_n$, $f_{\sigma_1,\ldots,\sigma_n}(\sigma_1 * \phi_1(x_1), \ldots, \sigma_n * \phi_n(x_n)) \leq f_{\sigma_1,\ldots,\sigma_n}(\sigma_1 * \phi_1(u_1), \ldots, \sigma_n * \phi_n(u_n))$ if and only if $f_{\tau_1,\ldots,\tau_n}(\tau_1 * \phi_1(x_1), \ldots, \tau_n * \phi_n(x_n)) \leq f_{\tau_1,\ldots,\tau_n}(\tau_1 * \phi_1(u_1), \ldots, \tau_n * \phi_n(u_n))$.

By considering families of functions rather than a single function for laws, Falmagne and Narens generalized the notion of 'dimensional constants' that appear in many laws. Their generalization allows for the formulation of behavioural laws (Falmagne and Narens 1983; Falmagne 1985) and physical laws (Falmagne 2004) that cannot be obtained by considering only a single function. Of course, Falmagne and Narens' theory also allows for the case of a single function, by allowing the family of functions to degenerate to a set consisting of a single function.

In many situations order meaningfulness is a testable condition, making it a preferable assumption to assuming a scale type for a dependent variable unless, of course, one already has a well-developed theory for the dependent variable.

In the Falmagne–Narens theory, the scale type of the dependent variable is not needed to obtain the law linking the independent and dependent variables.

For the case where the family $\mathscr{F}$ consists of a single function f of n- independent variables, Aczél, Roberts and Rosenbaum (1986) provided more general results. Through an insightful mathematical argument, they were able to characterize measurement laws using only measurability assumptions from real analysis about f instead of monotonicity or continuity assumptions. Aczél and Roberts (1989) use the general approach of Aczel et al. (1986) to derive measurement laws of economic interest.

Relation Between Meaningfulness and Invariance

Quantitative meaningfulness lacks a serious account as to why it is a good concept of meaningfulness; that is, it lacks a sound theory as to why it should yield correct results. Formulating a serious account for it is difficult. One tack (Krantz et al. 1971; Luce 1978; Narens 1981) is to observe that, if meaningfulness expresses valid qualitative relationships, then it must correspond to something purely qualitative, and therefore it should have a purely qualitative description. A long tradition in mathematics for formulating qualitative relationships that belong naturally to some structure or concept goes back to at least 19th-century geometry and was the centrepiece of the famous Erlanger Programme for geometry of Felix Klein. It was based on the idea that associated with each geometry was a set of transformations $\mathscr{T}$, and the relations and concepts belonging to the geometry were exactly those that were left invariant by all the transformations in $\mathscr{T}$. There are strong connections between (*a*) geometric techniques of establishing coordinate systems and measurement techniques for establishing scales, and (*b*) the Erlanger Programme's concept of 'geometric' and the measurement-theoretic concept 'meaningfulness'. To examine these connections, some definitions and conventions are needed.

Convention

Throughout the remainder of this article, it is assumed that $\mathscr{X}$ is a qualitative structure, which consists of a qualitative set X as its domain and relations based on X (called the *primitives of* $\mathscr{X}$); $\mathscr{N}$ is a numerically based structure, that is, $\mathscr{N}$ is a structure that has a subset of the real numbers as its domain; and $\mathscr{S}$ is the measurement scale consisting of all isomorphisms from $\mathscr{X}$ onto $\mathscr{N}$. (See measurement, theory of for a more detailed description of this kind of measurement scale.)

Qualitative Meaningfulness

An isomorphism of $\mathscr{X}$ onto itself is called a *symmetry* (or *automorphism)* of $\mathscr{X}$. It easily follows that if α is a symmetry of $\mathscr{X}$ and φ and ψ are elements of $\mathscr{S}$, then

- $\varphi^*\alpha$ is in $\mathscr{S}$,
- $\varphi^{-1*}\psi$ is a symmetry of $_X$,
- $\theta = \phi^*\psi^{-1}$ is an admissible transformation of $\mathscr{S}$, that is, $\varphi^*\eta$ is in $\mathscr{S}$ for each η in $\mathscr{S}$,and all admissible transformations can be obtained in the just mentioned manner by appropriate selections of φ and ψ.

An n-ary relation R on X is said to be *qualitatively meaningful* if and only if it is invariant under the symmetries of $\mathscr{X}$, that is, if and only if for each symmetry α of $\mathscr{X}$ and each $x_1,\ldots, x_n$ in X,

$$R(x_1,\ldots,x_n) \text{ iff } R(\alpha(x_1),\ldots,\alpha(x_n)).$$

Quantitative Meaningfulness

Although a relation T being 'quantitatively meaningful' was previously defined, it is defined again here to make explicit the role the scale $\mathscr{S}$ plays in qualitative meaningfulness: an n-ary relation T on N is said to be *quantitatively* $\mathscr{S}$*-meaningful* if and only if for each admissible transformation τ of $\mathscr{S}$ and each $r_1,\ldots, r_n$ in N,

$$T(r_1,\ldots,r_n) \text{ iff } T(\tau(r_1),\ldots,\tau(r_n)).$$

$\mathscr{S}$ can be used to interpret T as a relation U on X as follows. The n-ary relation U on X is said to be the $\mathscr{S}$*-inpt* of T if and only if for all φ in $\mathscr{S}$ and all $r_1,\ldots, r_n$

M

$$T(r_1,\ldots,r_n) \text{ iff } U\left(\varphi^{-1}(r_1),\ldots,\varphi^{-1}(r_n)\right).$$

Basic Result

The above definitions and relationships between symmetries and admissible transformations immediately yield the following theorem relating qualitatively and quantitatively meaningful relations:

Theorem *A relation T is quantitatively $\mathscr{S}$-meaningful if and only if its $\mathscr{S}$-inpt is qualitatively meaningful.*

The above theorem shows that each quantitatively meaningful relation has, through measurement, a corresponding qualitatively meaningful relation. Luce (1978) used this idea to provide a qualitative theory for practice of dimensional analysis in physics: Luce produced a qualitative structure $\mathscr{X}$ for measuring physical attributes. He showed that, under measurement, the quantitatively meaningful relationships among the attributes were the 'dimensionally invariant functions' of dimensional analysis. It is a principle of dimensional analysis that physical laws are such dimensionally invariant functions. Thus, by the just mentioned theorem, it then follows from the principles of dimensional analysis that each physical law corresponds to a qualitatively meaningful relation of $\mathscr{X}$. (Measurement-theoretic foundations for dimensional analysis can be found in Krantz et al. 1971; Luce et al. 1990; Narens 2002.)

Qualitative meaningfulness is just the Erlanger concept of 'geometric' applied to science. Mathematically, the two concepts are identical. The Erlanger Programme, as formulated by Klein (1872) and as used in mathematics, lacks a serious justification for assuming that the invariance of a relation under the symmetries of a geometry implies that the relation belongs to the geometry.

Scientific Definability

Narens (2002, 2007) sought to find a justification for Klein's assumption. He thought that a reasonable concept of a relation R belonging to a structure $\mathscr{X}$ was that R should somehow be definable in terms of the primitives of $\mathscr{X}$. But the usual concepts of 'definable' used in logic failed to provide a match with the Erlanger's concept of 'geometric'. Narens developed a new definability concept to capture the Erlanger Programme's concept of 'geometric'. He called the new concept *scientific definability.*

Scientific definability assumes that the quantitative world is constructed from relationships based on real numbers and is completely separated from the qualitative situation under investigation, $\mathscr{X}$, which is conceptualized as a qualitative structure. Unlike definability concepts from logic, scientific definability allows the free use of concepts from the quantitative world for defining relationships based on the domain X of a qualitative structure $\mathscr{X}$. Narens shows that a relation on X is qualitatively meaningful if and only if it is scientifically defined in terms of $\mathscr{X}$.

There is one obvious case where the Erlanger Programme appears to produce a rmkably poor concept of 'geometric'. This is where the geometry $\mathscr{X}$ has the identity function as its only symmetry, yielding that every relation on X is 'geometric', and for measurement situations where the scale consists of a single representation, making each relation on the domain of the numerical representing structure quantitatively meaningful, and thus, by the above theorem, each relation on X qualitatively meaningful. There are many important examples of this case, for example the geometry of physical universe under Einstein's general theory of relativity.

Narens (2002) provides generalizations of 'scientific definability' that appear to yield reasonable and productive concepts of 'geometric' ('qualitatively meaningful') for situations where the geometry (qualitative structure) has the identity as its only symmetry. The main idea for the generalizations is the following. Instead of formulating meaningfulness in terms of a single qualitative structure, a family $\mathscr{F}$ of isomorphic qualitative structures is used. It is assumed that all the structures in $\mathscr{F}$ have the same domain called the *common domain* (of $\mathscr{F}$). A relation R on the common domain is said to be *$\mathscr{F}$-meaningful* if and only if there exist a structure $\mathscr{X}$ in $\mathscr{F}$, primitives $R_{j_1},\ldots,R_{j_n}$ of $\mathscr{X}$, and a formula ϕ used for scientific definitions such that

(i) R has a scientific definition in terms $R_{j_1}, \ldots, R_{j_n}$ and ϕ, and

(ii) R has the same scientific definition for all $\mathscr{X}' = \left\langle X, R'_j \right\rangle_{j \in J}$ in $\mathscr{F}$; that is, R has the same scientific definition as in (*i*) but with $R_{j_1}, \ldots, R_{j_n}$ replaced by $R'_{j_1}, \ldots, R'_{j_n}$

For the case where $\mathscr{F}$ consists of a single structure, $\mathscr{F}$-meaningfulness coincides with qualitative meaningfulness.

See Also

► Measurement, Theory of

Bibliography

Aczél, J., and F.S. Roberts. 1989. On the possible merging functions. *Mathematical Social Sciences* 17: 205–243.

Aczél, J., F.S. Roberts, and Z. Rosenbaum. 1986. On scientific laws without dimensional constants. *Journal of Mathematical Analysis and Applications* 119: 389–416.

Falmagne, J.-C. 1985. *Elements of psychophysical theory.* New York: Cambridge University Press.

Falmagne, J.-C. 2004. Meaningfulness and order-invariance: Two fundamental principles for scientific laws. *Foundations of Physics* 34: 1341–1348.

Falmagne, J.-C., and L. Narens. 1983. Scales and meaningfulness of quantitative laws. *Synthese* 55: 287–325.

Klein, F. 1872. *Vergleichende Betrachtungen über neuere geometrische Vorschungen: Programm zu Eintritt in die philosophische Facultät und den Senat der Universität zu Erlangen*. Erlangen: Deichert.

Krantz, D.H., R.D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement*. Vol. 1. New York: Academic Press.

Luce, R.D. 1959. On the possible psychophysical laws. *Psychological Review* 66 (2): 81–95.

Luce, R.D. 1964. A generalization of a theorem of dimensional analysis. *Journal of Mathematical Psychology* 1: 278–284.

Luce, R.D. 1978. Dimensionally invariant numerical laws correspond to meaningful qualitative relations. *Philosophy of Science* 45: 1–16.

Luce, R.D., D.H. Krantz, P. Suppes, and A. Tversky. 1990. *Foundations of measurement*. Vol. 3. New York: Academic.

Narens, L. 1981. A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision* 13: 1–70.

Narens, L. 1985. *Abstract measurement theory.* Cambridge, MA: MIT Press.

Narens, L. 2002. *Theories of meaningfulness*. Mahwah: Lawrence Erlbaum and Associates.

Narens, L. 2007. *Introduction to the theories of measurement and meaningfulness and the role of invariance in science*. Mahwah: Lawrence Erlbaum and Associates.

Narens, L., and R.D. Luce. 1983. How we may have been misled into believing in the interpersonal comparability of utility. *Theory and Decision* 15: 247–260.

Pfanzagl, J. 1968. *Theory of measurement*. New York: Wiley. 2nd ed. Vienna: Physica, 1971.

Pickering, J.F., J.A. Harrison, and C.D. Cohen. 1973. Identification and measurement of consumer confidence: Methodology and some preliminary results. *Journal of the Royal Statistical Society Series A* 136: 43–63.

Roberts, F.S. 1980. On Luce's theory of meaningfulness. *Philosophy of Science* 47: 424–433.

Roberts, F.S. 1985. Applications of the theory of meaningfulness to psychology. *Journal of Mathematical Psychology* 229: 311–332.

Roberts, F.S., and C.H. Franke. 1976. On the theory of uniqueness in measurement. *Journal of Mathematical Psychology* 14: 211–218.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science* 103: 677–680.

Stevens, S.S. 1951. Mathematics, measurement and psychophysics. In *Handbook of experimental psychology*, ed. S.S. Stevens. New York: Wiley.

Suppes, P., and J.L. Zinnes. 1963. Basic measurement theory. In *Handbook of mathematical psychology*, ed. R.D. Luce, R.R. Bush, and E. Galanter, vol. 1. New York: Wiley.

Means, Gardiner Coit (1896–1988)

Leon H. Keyserling

Keywords

Berle, A. A.; Concentration; Corporations; Inflation; Inflation–unemployment trade-off; Means, G. C.; New deal; Ownership and control

JEL Classifications

B31

The Modern Corporation and Private Property appeared in 1932, co-authored by Means and Adolph Berle. This book fused the abilities of a great economist and a great lawyer, and became deservedly famous.

The prevalent economic and legal thinking at that time did not recognize adequately the emergence of corporate giantism. It envisaged a system characterized in the main by small private enterprises. And it assumed that this worked well because the law of supply and demand would determine price levels and thus automatically produce adjustments assuring the greatest good of the greatest number. This laissez-faire approach made private property almost sacrosanct and almost free from public intervention.

Berle and Means *proved* that this was not how the economy actually worked. As they revealed, the monster size of existing corporations and their dominating power negated the attributes of private property as then conceived; their ability and determination to fix or 'administer' prices prevented the benign operation of supply and demand. These findings suggested that increased government intervention in the private sector was essential in the public interest; and this was fortified by the book's additional findings *in re* economic concentration and the separation between ownership and control. Thus the book indicated among other things the need for legal change, including judicial reinterpretations of governmental powers under the Constitution.

Substantial parts of the book were used to support New Deal and judicial action between 1933 and 1939, which viewed corporations and private property in a new light. And the New Deal brought Means to Washington. He alone among three economic advisers to the Secretary of Agriculture dealt with the effect of farm conditions upon the overall economy. Next, he was Director of the industrial division of the National Resources Planning Board (NRPB), where he developed techniques for depicting what composition of business activity would maintain full employment. This type of work, continued by him on the staff of the Committee for Economic Development (after a spell as fiscal analyst in the Bureau of the Budget) was intrinsic to the Committee's portrayal of the post-war markets requisite to full employment.

During subsequent decades, Means poured forth a Niagara of writing and speeches. Insistently, he built upon his original thesis of corporate power, especially through its pricing practices. Refuting the prevalent view among economists that there is a 'trade-off' between unemployment and inflation, he showed that the great increases in inflation during recent decades have come mainly, not during a highly used economy near full employment, but rather during periods when the economy moved into stagnation and recession. This squared with his early finding that the modern corporation can and does lift prices to compensate for low volume. It also revealed that his humanistic concern about full employment and economic justice must reject the frequent and unsuccessful efforts to achieve price stability by spawning the misery of vast unemployment. Always, unlike so many economists, Means eschewed outmoded or untested theories, and spent himself in exhaustive empirical studies in aid of his own analysis and policy recommendations.

See Also

► Berle, Adolf Augustus, Jr. (1895–1971)

Selected Work

1932. (With Berle A. A). *The modern corporation and private property.* New York: Commerce Clearing House.

Mean-Variance Analysis

Harry M. Markowitz

Abstract

Mean-variance analysis is concerned with combining risky assets in a way that minimizes the variance of risk at any desired mean return. In the use of mean-variance analysis for actual money management, the issue is how to

estimate the large number of required covariances. Many-factor models of covariance are widely used, as are scenario and combined scenario and factor models, and constant correlation models. This simplifies the parameter estimation problem and can accelerate the computation of efficient sets for analyses containing hundreds of securities.

Keywords

Capital asset pricing models; Constant correlation models; Corner portfolio; Covariance; Dynamic programming; Factor models; Liquidity preference; Markowitz, H. M.; Mean-variance analysis; Option Prices; Quadratic Approximation; Risk aversion; Scenario models

JEL Classifications

G1

In a mean–variance portfolio analysis (Markowitz 1959) an n-component vector (portfolio) X is called feasible if it satisfies

$$AX = b\, X \geq O$$

where A is an $m \times n$ matrix of constraint coefficients, and b an m-component constant vector. An EV combination is called feasible if

$$E = \mu^T X$$

$$V = X^T C X$$

for some feasible portfolio. Here E is the expected return of the portfolio, V the variance of the portfolio, μ the vector of expected returns of securities, and C a positive semidefinite covariance matrix of returns among securities.

A feasible EV combination is called inefficient if some other feasible combination has either less V and no less E, or else greater E and no greater V. A feasible EV combination is called efficient if it is not inefficient. A feasible portfolio X is efficient or inefficient according to whether its EV combination meets the one definition or the other. As in linear programming, the constraints ($AX = b$, $X \geq O$) can represent inequalities by introducing slack variables, and can incorporate variables which are allowed to be negative, by separating the positive and negative parts of such variables.

Markowitz (1956) shows that if V is strictly convex over the set of feasible portfolios – for example when C is positive definite – the set of efficient portfolios is piecewise linear, and the set of efficient EV combinations is piecewise parabolic. There may or may not be a kink in the efficient EV set at a 'corner portfolio', where two pieces of the efficient portfolio set meet. Markowitz (1959, Appendix A), shows for arbitrary semidefinite C that, while there may be more than one efficient portfolio for given efficient EV combination, there is a piecewise linear set of efficient portfolios which contains one and only one efficient portfolio for each efficient EV combination. The piecewise linear nature of the efficient set is illustrated graphically, for small n, in Markowitz (1952, 1959).

The fact that the mean-variance analysis selects a portfolio for only one period does not imply that the investor plans to retire at the end of the period. Rather, it assumes that in the dynamic programming (Bellman 1957) solution to the many period investment problem, current wealth is the only state variable to enter the implied single period utility function (see Markowitz 1959, Ch. 13; Samuelson 1969; Ziemba and Vickson 1975). Mossin (1968) shows conditions under which the optimum solution to the many period problem is 'myopic' in that the single period utility function is the same as an end-of-game utility function. This is an example of – but not the only example of – a class of games in which wealth is the only state variable.

The Markowitz (1959) justification for the use of mean-variance analysis further assumes that if one knows the E and V of a portfolio one can estimate with acceptable accuracy the expected value of the one-period utility function. Samuelson (1970) and Ohlson (1975) present conditions under which mean and variance are asymptotically sufficient as the length of holding periods – that is, the intervals between portfolio revisions – approaches zero. For 'long'

holding periods, for example for time between revisions as long as a year, Markowitz (1959), Young and Trent (1969), Levy and Markowitz (1979), Pulley (1981) and Kroll et al. (1984) have each found mean-variance approximations to be quite accurate for a variety of utility functions and historical distributions of portfolio return.

This leads to an apparent anomaly: if you know mean and variance you practically know expected utility; the mean-variance approximation to expected utility is based on a quadratic approximation to the single-period utility function; yet Arrow (1965) and Pratt (1964) show that any quadratic utility function has the objectionable property that an investor with such a utility function becomes increasingly averse to risks of a given dollar amount as his wealth increases. Levy and Markowitz (1979) show that the anomaly disappears if you distinguish three types of quadratic approximation:

1. Assuming that the investor has a utility-of-wealth function that remains constant through time – so that as the investor's wealth changes he moves along the curve to a new position – fit a quadratic to this curve at some instant of time, and continue to use this same approximation subsequently. (Note that the assumption here, that the investor has a constant utility-of-wealth function is sufficient, but not necessary, for the investor to have a single period utility function at each period.)
2. Fit the quadratic to the investor's current single period utility function. For example, if the investor has an unchanging utility-of-wealth function, choose a quadratic to fit well near current wealth (i.e. near portfolio return equal zero).
3. Allow the quadratic approximation to vary from one portfolio to another, that is, let the approximation depend on the mean, and perhaps the standard deviation, of the probability distribution whose expected value is to be estimated.

The Pratt-Arrow objections apply to an approximation of type (1). The approximations proposed in Markowitz (1959) are of types (2) and (3). Levy and Markowitz (1979) show that, under quite general assumptions, the type 3 mean- variance maximizer has the same risk aversion in the small (in the sense of Pratt) as does the original expected utility maximizer.

Uses of Mean-Variance Analysis

Two areas of use deal with: (a) actual portfolio management using mean–variance analysis, and (b) implications for the economy as a whole of the assumption that all investors act according to the mean–variance criteria. We refer to these, respectively, as 'normative' and 'positive' uses of mean–variance analysis.

The positive application of mean-variance analysis is dealt with elsewhere in this Dictionary. Seminal works in the field include the Tobin (1958) analysis of liquidity preference; and the Sharpe (1964), Lintner (1965) and Mossin (1966) Capital Asset Pricing Models (CAPMs). As in the Tobin model, these CAPMs assume that the investor can either lend all he has or borrow all he wants at the same 'risk-free' rate of interest. From this assumption (plus assumptions that all investors have the same beliefs and seek mean–variance efficiency subject to the same constraint set) they conclude that the excess return on each security (its expected return minus the risk-free rate) is proportional to its 'beta', where the latter is the regression of the security's return against the return of the market as a whole. Black (1972) drops the assumption that the investor can borrow at a risk-free rate; assumes instead that the investor can sell short and use the proceeds to buy long; and derives a formula for excess return just like that of Sharpe–Lintner–Mossin except that the expected return on a zero-beta portfolio is substituted for the risk-free rate in the formula for excess return. Merton (1969) has developed mean–variance theory in continuous time. This has been used, for example, in the analysis of option prices by Black and Scholes (1973) from which a vast literature of further implications followed.

As compared with the models used in normative analysis, the models of positive analysis tend

to use quite simple constraint sets and other special assumptions (e.g. all investors have the same beliefs). The justification for such assumptions is that they give concrete, therefore testable, implications; and indeed have been the subject of extensive empirical testing.

In the use of mean-variance analysis for actual money management, the question immediately arises as to how to estimate the large number of required covariances. Sharpe (1963) concluded, and Cohen and Pogue (1967) confirmed, that a simple one-factor model of covariance was sufficient. King (1966) showed that, in addition to one pervasive factor, there were ample industry sources of covariance. By the mid-1970s it was clear to many practitioners that the one-factor model was not adequate, since, for example, sometimes 'the market', as measured by some broad index, went up while high beta stocks went down, to an extent that could not be explained by chance. Many-factor models such as that of Rosenberg (1974) are now widely used.

Other models of covariance used in practice include scenario and combined scenario and factor models (Markowitz and Perold 1981), and a model which assumes that all correlation coefficients are the same (Elton and Gruber 1973). The use of factor, scenario or constant correlation models, in addition to simplifying the parameter estimation problem, can considerably accelerate the computation of efficient sets for analyses containing hundreds of securities. For example, the Perold (1984) code will solve large portfolio selection problems for arbitrary A and C, but is especially efficient in handling upper bounds on variables and sparse (mostly zero) A and C matrices. (The introduction of 'dummy' securities into the analysis allows one to 'sparsify' the C matrix for factor, scenario or constant correlation models.) Even faster solutions are obtained by Elton et al. (1976, 1978) for the one-factor and constant correlation models for certain common constraint sets.

See Also

- ▶ Capital Asset Pricing Model
- ▶ Finance

Bibliography

Arrow, K. 1965. *Aspects of the theory of risk bearing*. Helsinki: Yrjö Jahnsson Foundation.

Bellman, R.E. 1957. *Dynamic programming*. Princeton: Princeton University Press.

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

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Cohen, J.K., and J.A. Pogue. 1967. An empirical evaluation of alternative portfolio-selection models. *Journal of Business* 40: 166–193.

Elton, E.J., and M.J. Gruber. 1973. Estimating the dependence structure of share prices. *Journal of Finance* 28: 1203–1232.

Elton, E.J., M.J. Gruber, and M.W. Padberg. 1976. Simple criteria for optimal portfolio selection. *Journal of Finance* 31: 1341–1357.

Elton, E.J., M.J. Gruber, and M.W. Padberg. 1978. Simple criteria for optimal portfolio selection: Tracing out the efficient frontier. *Journal of Finance* 33: 296–302.

King, B.F. 1966. Market and industry factors in stock price behavior. *Journal of Business* 39: 139–190.

Kroll, Y., H. Levy, and H.M. Markowitz. 1984. Mean variance versus direct utility maximization. *Journal of Finance* 39: 47–61.

Levy, H., and H.M. Markowitz. 1979. Approximating expected utility by a function of mean and variance. *American Economic Review* 69: 308–317.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.

Markowitz, H.M. 1956. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly* 3: 111–133.

Markowitz, H.M. 1959. *Portfolio selection: Efficient diversification of investments*. New Haven: Yale University Press. Reprinted, New York: Wiley, 1970.

Markowitz, H.M., and A. Perold. 1981. Portfolio analysis with factors and scenarios. *Journal of Finance* 36: 871–877.

Merton, R.C. 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics* 51 (3): 247–257.

Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34: 768–783.

Mossin, J. 1968. Optimal multiperiod portfolio policies. *Journal of Business* 41: 215–229.

Ohlson, J.A. 1975. The asymptotic validity of quadratic utility as the trading interval approaches zero. In *Stochastic optimization models in finance*, ed. W.T. Ziemba and R.G. Vickson. New York: Academic Press.

Perold, A.F. 1984. Large-scale portfolio optimization. *Management Science* 30: 1143–1160.

Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32: 122–136.
Pulley, L.B. 1981. A general mean-variance approximation to expected utility for short holding periods. *Journal of Financial and Quantitative Analysis* 16: 361–373.
Rosenberg, B. 1974. Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis* 9 (2): 263–274.
Samuelson, P.A. 1969. Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics* 51 (3): 239–246.
Samuelson, P.A. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances and higher moments. *Review of Economic Studies* 37: 537–542.
Sharpe, W.F. 1963. A simplified model for portfolio analysis. *Management Science* 9 (2): 277–293.
Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25: 65–86.
Young, W.E., and R.H. Trent. 1969. Geometric mean approximations of individual security and portfolio performance. *Journal of Financial and Quantitative Analysis* 4 (2): 179–199.
Ziemba, W.T., and R.G. Vickson, eds. 1975. *Stochastic optimization models in finance*. New York: Academic Press.

Measure Theory

A. P. Kirman

JEL Classifications
C0

Measure theory is that part of mathematics which is concerned with the attribution of weights of 'measure' to the subsets of some given set. Such a measure is required to satisfy a natural condition of additivity, that is that the measure of the union of disjoint sets should be equal to the sum of the measure of those sets. The fundamental problems of measure arise when one has to treat infinite sets or infinite unions of sets. It is perhaps not clear why such a tool should be of use in economics.

Apart from the rather trivial observation that, since measure theory provides the basis for probability theory it underlies all of the economics of uncertainty, there have been direct applications of this theory to several basic problems in economic theory (for a more detailed account, see Kirman 1982). A first example of such an application is given by the idea of 'pure' or 'perfect' competition. The fundamental characteristic of a perfectly competitive economic situation is one in which no individual can influence the outcome. Thus, in a competitive market economy, although prices are the result of the collective activity of all the agents, no individual by acting alone can modify them and hence takes them as given. Now strictly speaking in a finite economy this cannot be true and in the work of Torrens, Cournot and Edgeworth can be found lengthy discussions as to whether it is rational for individuals to behave in a perfectly competitive way. Indeed, as Viner once observed, the fact that it is profitable for them to do otherwise is a 'skeleton in the cupboard of free trade'.

Economists have typically avoided the contradiction involved in analysing economies in which individuals do have positive influence but behave as if they do not, by saying that individuals behave 'as if' or 'believe that' they have no effect on the outcome. To a mathematician there is no contradiction involved in the idea of individual elements having no weight but sets of such elements having positive weight. If we think of the unit interval, each point has no length but sub-intervals made up of such points do have positive weight. This is, of course, due to the fact that there are infinitely many, indeed a continuum, of such points. Aumann (1964) in his path-breaking article made use of these ideas to define an 'ideal' or 'continuum' economy which corresponded logically to the idea of perfect competition. If instead of the set of agents A in an economy being finite, we substitute the unit interval [0, 1] a continuum exchange economy can be defined by $e[0,1] \to \mathscr{P}_{\text{mo}} \times R^l_+$ where l is the number of goods and $\mathscr{P}_{\text{mo}}$ is the set of monotonic continuous preferences on R^l_+ positive orthant of Euclidian l space. Thus with each agent

or point is associated a preference relation and an initial bundle of goods. Now we have defined an economy which has the right framework for perfect competition. To be able to use this model requires a little more. If we think of an allocation f which assigns to each agent a bundle of goods how do we say that what is allocated is equal to the sum of the initial resources $e(a)$ of the agents. To write

$$\sum_{\alpha \in A} e(a) = \sum_{\alpha \in A} f(a)$$

no longer makes sense. However, in a finite economy with n agents, we could also write.

$$\underset{\text{average allocation}}{1/n \sum e(a)} \quad = \quad \underset{\text{averages resources}}{1/n \sum f(a)}$$

without changing anything. In the continuum economy, just such a statement can be made by writing

$$\int e(a) = \int f(a).$$

When taking an average in this way by integrating we are assigning weights to the various subsets of agents. In other words, we integrate 'with respect to some measure μ'. In the case where $A = [0, 1]$ there is a natural measure (Lebesgue measure) which corresponds to the length of the intervals which make up a set. This allows us to carry through all the standard analysis in such an economy and indeed allows one to obtain two interesting results which do not hold in finite economies. The first is that in such an economy a competitive equilibrium exists even if preferences are not convex. The second is that the set of Walrasian allocations $W(\varepsilon)$ is equal to the set of allocations which no coalition can improve upon, called the core $C(\varepsilon)$ of the economy (see cores). This last result is the 'perfect' or 'ideal' version of an old result of Edgeworth. In fact, Aumann's results can be shown to be approximately true for large but finite economies and thus, as one might hope, the ideal case gives us a good idea of what happens in large economies.

Two observations are in order. In fact, the choice of the unit interval and Lebesgue measure is arbitrary. All that one needs is a *measure space* $(A, \mathscr{A}, \mu)$ where A is the set of agents $\mathscr{A}$ is the collection of subsets or coalitions of agents and μ is the measure of these subsets. $\mathscr{A}$ can be thought of as the set of all subsets of A although strictly speaking this is not correct. What is required to model perfect competition is that no individual has weight. Thus one must add the condition that the measure space be 'atomless' that is for any set C with $\mu(C) > 0$ there must be a subset B contained in C with $\mu(C) > \mu(B) > 0$. This is in contradiction with the standard term 'atomistic competition' which is supposed to describe perfect competition. Another aspect of economies which implicitly makes use of the notion of a continuum economy is the discussion of the *distribution of agents' characteristics*. It is common practice in economics to use a continuous function such as the Pareto distribution to describe the income distribution. For this to be fully appropriate a continuum economy is needed. How may we formally describe distributions? Suppose that we start with a measure space of agents as explained above. Now consider an economy ε i.e. an attribution of preferences and initial resources to each agent. Take a set B in the characteristics space and consider the set C of those agents who have characteristics in B i.e. $C = \varepsilon^{-1}(B)$ Now let the measure $\psi(B) = \mu(C)$ Thus the measure μ on the set of agents induces another measure ψ on the set of characteristics. This defines the distribution of characteristics in that economy. Now, one could maintain that a good argument for the distribution approach would be that two economies with the same distribution of characteristics should have the same equilibria, for example. Hildenbrand (1975) gives a detailed discussion of this problem. The general merit of the distribution approach is of course that individualistic descriptions of the characteristics of agents make little economic sense in very large economies. Furthermore, in such economies, putting conditions on the distribution of characteristics may help in restricting the class of outcomes that may be observed.

An illustration of the necessity for this is given by the results of Sonnenschein (1973) and Debreu (1974) which show that all the standard individualistic assumptions on individuals put no restrictions on the aggregate excess demand of an economy other than continuity and Walras's Law. This means that there is essentially no *a priori* restriction on the form of aggregate excess demand functions and hence on observable outcomes. Indeed, in finite economies, even specifying the income distribution does not help (see Kirman and Koch 1986). However, Hildenbrand (1983) has shown that, in a continuum economy, if one puts a condition on the income distribution, then the 'law of demand' is satisfied. This 'normality' of goods with respect to prices is a fairly strong restriction on excess demand functions and indicates that other results in the same direction might be obtained.

Rather than make assumptions about the specific form of the distribution it is sometimes useful to be able to say something about how 'dispersed' agents characteristics are. This involves requiring that the 'support' of the measure ψ representing the distribution of characteristics, that is the smallest set which has full measure, should not be 'too small'. For example, a bothersome feature of the standard assumption of convex rather than strictly convex preferences is that demand is a 'correspondence' rather than a function. This involves considerable technical difficulties. However, it has been shown by various authors (an account may be found in Mas-Colell (1985) for example) that if the support of the distribution of agents characteristics is sufficiently large then aggregate demand will be a function and not a correspondence.

Another use of measure theory is to give precision to the idea that phenomena are 'unlikely'. Thus one cannot exclude, for example, the possibility that an economy will have an infinite set, even a continuum, of equilibrium allocations. However, as Debreu (1970) has shown, 'almost no' economies have this property. To see the idea consider an Edgeworth box representing a two man, two good exchange economy. Each point in the box can be considered as a possible location of the individual endowments. Naturally, the equilibria vary with initial endowments. What is true is that if we consider the set of endowments which give rise to infinite equilibria, its 'area' or 'measure' is zero. Thus the probability that an economy drawn at random, in some sense, will have infinite equilibria is zero.

A classic problem which has received considerable attention is that of how to divide some object fairly among n individuals. Suppose that the object is not homogeneous, a cake with different layers for example, then if an individual assigns value 1 to the whole cake he can give a value to any piece of the cake. In other words, each individual i has a measure μ_i on the cake. It has been shown that it is possible to find partitions of U ($U_1 \ldots U_n$) so that each individual i considers that his share U_i is worth more than $1/n$ of the cake. This does not exclude some individuals being jealous of each other. However, Dubins and Spanier (1961) have shown that it is possible to find partitions where each individual considers that all the pieces U_i of the cake are worth $1/n$. Thus:

$$\mu(U_j) = 1/n \; i,j = 1 \ldots n$$

and everybody believes that the division is perfectly equitable.

Another illustration of the measure theoretic approach is the following. Arrow (1963) discussed the problem of establishing a rule which aggregates individual preferences on a set of social allocations into social preferences. He showed that if all individual preferences are allowed then no rule satisfying certain basic axioms exists. In particular, he showed that his first axioms implied that there must be a 'dictator', who has the property that if he prefers state x to state y, then society prefers x to y. Fishburn (1970) showed that this was not true in a society with an infinite number of individuals, thus raising hopes that in large economies Arrow's result loses its importance. In fact, this is not the case, Arrow's axioms impose a very special structure on those sets of individuals who 'dictate' society's preferences in the above sense. This structure implies that no matter how large the finite economy there

will always be a dictator. Thus the infinite case is exceptional. However, in the infinite society individuals make little sense and one can give a measure theoretic equivalent of Arrow's result. For a society in which the set of individuals is represented by the unit interval then any dictatorial set C contains a dictatorial set B with positive but smaller measure that is

$$\mu(C) > \mu(B) > 0 \text{ with } B \subset C.$$

Thus there are dictatorial sets of arbitrarily small measure.

As a last example consider the problem of 'temporary equilibrium'. In an economy in which one can only transfer wealth to the future by keeping money and in which individuals anticipate future prices, one wishes to find an equilibrium for the goods and money markets today. Each individual forms a distribution over tomorrow's prices p_2 as a function ψ of today's prices p_1. Now if for example, individuals always expect prices tomorrow to be higher than today there may be no incentive at any prices to hold money. In this case, there can be no equilibrium. However, if we require that prices tomorrow should not be 'too dependent' on today's then equilibrium exists. Formally, we require that the family of the price distributions ψ over all prices should be 'tight'. An explanation of this with results is given by Grandmont (1977). However, intuitively, it is clear that one excludes the ever increasing expectations that lead to hyperinflation.

These examples illustrate the ways in which a formal mathematical tool, measure theory, has been incorporated into economic theory. In particular, its use in characterizing ideal economies, those corresponding to the notion of perfect competition, has been invaluable.

See Also

- Cores
- Functional Analysis
- Lyapunov Functions
- Non-standard Analysis

Bibliography

Any standard text in measure theory such as Halmos (1971) will give the essential mathematical notions, and more specialized references are given in the bibliographies of the articles cited here.

Arrow, K.J. 1963. *Social choice and individual values*. 2nd ed. New York: Wiley.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1 (1): 15–21.

Dubins, I.E., and E.H. Spanier. 1961. How to cut a cake fairly. *American Mathematical Monthly* 1: 1–17.

Fishburn, P.C. 1970. Arrow's impossibility theorem, concise proof and infinite voters. *Journal of Economic Theory* 2: 103–106.

Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45 (3): 535–572.

Halmos, P.R. 1961. *Measure theory*. 7th ed. Princeton: Van Nostrand.

Hildenbrand, W. 1975. Distributions of agents characteristics. *Journal of Mathematical Economics* 2: 129–138.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51 (4): 997–1020.

Kirman, A.P. 1982. Chapter 5: Measure theory with applications to economics. In *Handbook of mathematical economics*, ed. K.J. Arrow and M. Intriligator, vol. 1, 159–209.

Kirman, A.P., and K.J. Koch. 1986. Market excess demand in economies with identical preferences and co-linear endowments. *Review of Economic Studies* 53 (3): 457–464.

Mas-Colell, A. 1985. *The theory of general economic equilibrium*. Cambridge: Cambridge University Press.

Sonnenschein, H. 1973. Do Walras's identity and continuity characterise the class of community excess demand functions. *Journal of Economic Theory* 6 (4): 345–354.

M

Measurement

Marcel Boumans

Abstract

Measurement theory takes measurement as the assignment of numbers to properties of an

empirical system so that a homomorphism between the system and a numerical system is established. To avoid operationalism, two approaches can be distinguished. In the axiomatic approach it is asserted that if the empirical system satisfies a certain set of axioms such a homomorphism can be constructed. In the empirical approach, empirical adequacy is established by aiming at accuracy, precision and standardization. Precision is achieved by least-squares-errors methods, accuracy by calibration and standardization by the involvement of independent theoretical and empirical studies.

Keywords

Axiomatic index theory; Axiomatic theory; Calibration; Ceteris paribus; Fisher, I.; Functional equation theory; Index numbers; Measurement error models; Measurement theory; Model theory of measurement; Operationalism; Passive observations; Price indexes; Representation theorems; Representational theory of measurement; Structural parameters; Uniqueness theorems

JEL Classifications

B4

The dominant measurement theory is the representational theory of measurement (RTM), which takes measurement as a process of assigning numbers to attributes of the empirical world in such a way that the relevant qualitative empirical relations among these attributes are reflected in the numbers themselves as well as in important properties of the number system.

The RTM defines measurement set-theoretically. Given a set of empirical relations $\mathbf{R} = \{R_1, \ldots, R_m\}$ on a set of extra-mathematical entities $\mathbf{X}$ and a set of numerical relations $\mathbf{P} = \{P_1, \ldots, P_m\}$ on the set of numbers $\mathbf{N}$ (in general a subset of the set of real numbers), a function φ from $\mathbf{X}$ into $\mathbf{N}$ takes each R_i into P_i, $i = 1, \ldots, m$; provided that the elements $x, y, \ldots$, in $\mathbf{X}$ stand in relation R_i if and only if the corresponding numbers $\varphi(x), \varphi(y), \ldots$; stand in relation P_i. In other words, measurement is conceived of as establishing homomorphisms (also called scales) from empirical relational structures $\langle \mathbf{X}, \mathbf{R} \rangle$ into numerical relational structures $\langle \mathbf{N}, \mathbf{P} \rangle$. A numerical relational structure representing an empirical relational structure is also called a model, therefore the RTM is sometimes called the model theory of measurement.

The problem is that when the requirements for choosing a representation or model are not further qualified, it can easily lead to an operationalist position, which is most explicitly expressed by Stevens (1959, p. 19): 'Measurement is the assignment of numerals to objects or events according to rule – any rule.' A model should meet certain criteria to be considered homomorphic to an empirical relational structure. In economics, there are two different foundational approaches, an axiomatic and an empirical approach (Boumans 2007).

Axiomatic Theory

The axiomatic theory is most comprehensively presented in Krantz et al. (1971–90). According to this literature the foundations of measurement are established by axiomatization. The analysis into the foundations of measurement involves, for any particular empirical relation structure, the formulation of a set of axioms that is sufficient to establish two types of theorems, a representation theorem and a uniqueness theorem.

A representation theorem asserts that if a given relational structure satisfies certain axioms, then a homomorphism into a certain numerical relational structure can be constructed. A uniqueness theorem sets forth the permissible transformations $\varphi \rightarrow \varphi'$. A transformation $\varphi \rightarrow \varphi'$ is permissible if and only if φ and φ' are both homomorphisms of $\langle \mathbf{X}, \mathbf{R} \rangle$ into the same numerical structure $\langle \mathbf{N}, \mathbf{P} \rangle$.

Probably the first example of the axiomatic approach in economics is Frisch (1926), in which three axioms define utility as a quantity. The work more often referred to as the one that introduced the axiomatic approach to economics, however, is Von Neumann and Morgenstern (1944). They required the transformation φ :

$\mathbf{X} \rightarrow \mathbf{N}$ to be order-preserving: $x > y$ implies $\varphi(x) > \varphi(y)$, and linear:

$$\varphi(\alpha x + (1-\alpha)y) = \alpha\varphi(x) + (1-\alpha)\varphi(y), where \quad \alpha \in (0,1).$$

Another field in economics in which the axiomatic approach has been influential is the axiomatic index theory. This theory originates from Fisher's work on index numbers (1911, 1922). Fisher evaluated in a systematic manner a very large number of indices with respect to a number of criteria. These criteria were called 'tests'. Fisher himself did not expect that it would be possible to devise an index number that would satisfy all these tests. Moreover, Frisch (1930) proved the impossibility of maintaining a certain set of Fisher's tests simultaneously. It is, however, Eichhorn and Voeller (1976) who provide a definite evaluation of Fisher's tests by their axiomatic approach.

Eichhorn and Voeller (1976) look systematically at the inconsistencies between various tests (and how to prove such inconsistencies) by means of the functional equation theory. Functional equation theory is transferred into index theory if the price index is defined as a positive function $P(\mathbf{p}_s, \mathbf{x}_s, \mathbf{p}_t, \mathbf{x}_t)$ that satisfies a number of axioms, where $\mathbf{p}$ is a price vector and x a commodity vector, and the subscripts are time indices. These axioms do not, however, determine a unique form of the price index function. Several additional tests are needed for assessing the quality of a potential price index. Both axioms and tests are formalized as functional equations. When the axioms are formalized as functional equations, inconsistency theorems can then be proven by showing that for the relevant combinations of functional equations, the solution space is empty.

In current axiomatic index theory, axioms specify mathematical properties that are essential or desirable for an index formula. One of the problems of axiomatic index theory is the impossibility of simultaneously satisfying all axioms. In practice, however, a universally applicable solution to this problem is not necessary. The specifics of the problem at hand, including the purpose of the index and the characteristics of the data, determine the relative merits of the possible attributes of the index formula.

Empirical Approach

Relation-rich structures, in contrast to object-rich structures, do not lend themselves to axiomatization. This does not mean, however, that measurement is impossible, but that a representation should, beside theoretical requirements, also satisfy empirical criteria. Moreover, economic measurements are often developed for purposes of economic policy; so, representations should also satisfy criteria of applicability. For example, a national account system should be a consistent structure of interdependent definitions, enabling uniform analysis and comparison of various economic phenomena.

To understand empirical measurement approaches, let us consider the problem of measuring a property x of an economic phenomenon. $y_i (i = 1, \ldots, n)$ are repeated observations to be used to determine value x. Each observation involves an observational error, ε_i. This error term, representing noise, reflects the operation of many different, sometimes unknown, background conditions, indicated by B:

$$y_i = f(x, B_i) = f(x, 0) + \varepsilon_i (i = 1, \ldots n) \quad (1)$$

Now, accuracy is obtained by reducing noise as much as possible. One way of obtaining accuracy is by taking care that the background conditions B are held constant, in other words, that *ceteris paribus* conditions are imposed. To show this, Eq. (1) is rewritten to expresses how x and possible other conditions (B) influence the observations:

$$\Delta y = f_x \Delta x + f_B \Delta B = f_x \Delta x + \Delta\varepsilon \quad (2)$$

Then, imposing *ceteris paribus* conditions: $\Delta B \approx 0$ reduces noise.

However, *ceteris paribus* conditions imply full control of the circumstances and complete knowledge of all potential influence quantities.

However, in economics we have often to deal with open systems, in which full control is not feasible. As a result, accuracy has to be obtained by modelling in a specific way. To measure x a model M has to be specified of which the values of the observations y_i functions as input and the output estimate $\hat{x}$ as measurement result: $\hat{x} = M[y_i; \alpha]$, where α denotes the parameter set of the model. If one substitute Eq. (1) into model M, one can derive that, assuming that M is a linear operator (usually the case):

$$\hat{x} = M[f(x) + \varepsilon; \alpha] = M_x[x; \alpha] + M_\varepsilon[\varepsilon; \alpha]. \quad (3)$$

A necessary condition for the measurement of x is that a model M must entail a representation of the measurand, M_x, and a representation of the environment of the measurand, M_ε.

The performance of a model built for measuring purposes is described by the terms accuracy and precision. In metrology, accuracy is defined as the statement about the closeness of the model's outcome to a value declared as the standard. Precision is a statement about the spread of the estimated measurement errors. The usual procedure to attain precision is by minimizing the variance of errors. The procedure to obtain accuracy is calibration, which is the establishment of the relationship between values indicated by a model and the corresponding values realized by standards. So, we can split the measurement error in three parts:

$$\hat{\varepsilon} = \hat{x} - x = M_\varepsilon + (M_x = S) + (S - x) \quad (4)$$

where S represents a standard value. The error term M_ε is reduced as much as possible by aiming at precision. (M_x - S) is the part of the error term that is reduced by calibration. The reduction of the last term (S - x) is called standardization and is dealt with by finding an invariant structure representing the measurement system.

Attempting to find these invariant structures, we have to deal with the so-called problem of passive observation: it is not possible to identify the reason for a disturbing influence, say z, being negligible, $f_z\Delta z \approx 0$. We cannot distinguish whether its potential influence is very small, $f_z \approx 0$, or whether the factual variation of this quantity over the period under consideration is too small, $\Delta z \approx 0$. The variation of z is determined by other relationships within the economic system. In some cases, a virtually dormant quantity may become active because of changes in the economic system elsewhere. Each found empirical relationship is a representation of a specific data-set. So, for each data-set it is not clear whether potential influences are negligible or only dormant. This is what Haavelmo (1944) called the problem of autonomy. Some of the empirical found relations have very little 'autonomy' because their existence depends upon the simultaneous fulfilment of a great many other relations. Autonomous relations are those relations that could be expected to have a great degree of invariance with respect to various changes in the economic system.

This problem of autonomy is dealt with by the following modelling strategy: when a relationship appears to be inaccurate, this is an indication that a potential factor is omitted. As long as the resulting relationship is inaccurate, potential relevant factors should be added. The expectation is that this strategy will result in the fulfilment of two requirements: (*a*) the resulting model captures a complete list of factors that exert large and systematic influences and (*b*) all remaining influences can be treated as a small noise component. The problem of passive observations is solved by accumulation of data-sets: the expectation is that we converge bit by bit to a closer approximation to the complete model, as all the most important factors reveal their influence. This strategy, however, is not applicable in cases when there are influences that we cannot measure, proxy or control for, but which exert a large and systematic influence on the outcomes.

A very influential paper in macroeconometrics (Lucas 1976) showed that the estimated so-called structural parameters (α) achieved by the above strategy are not invariant under changes of policy rules. Policy-invariant parameters should be obtained in an alternative way. Either they could be supplied from independent microeconometric studies, accounting identities or institutional facts, or they are chosen to secure a good match between a selected set of characteristics of the

actual observed time series and those of the simulated model output. These alternative ways of obtaining parameter values are all covered by the label calibration. It is important that, whatever the source, the facts being used for calibration should be as stable as possible. An important result of this calibration strategy is that for accurate measurement it is no longer required for representations to be homomorphic to an empirical relational structure.

See Also

- ▶ Calibration
- ▶ Ceteris Paribus
- ▶ Econometrics
- ▶ Meaningfulness and Invariance
- ▶ Measurement Error Models
- ▶ Measurement, Theory of

Bibliography

Boumans, M. 2007. *Measurement in economics: A handbook*. Elsevier.

Eichhorn, W., and J. Voeller. 1976. *Theory of the price index*. Berlin: Springer.

Fisher, I. 1911. *The purchasing power of money*, 1963. New York: Kelley.

Fisher, I. 1922. *The making of index number*, 1967. New York: Kelley.

Frisch, R. 1930. Necessary and sufficient conditions regarding the form of an index number which shall meet certain of Fisher's tests. *Journal of the American Statistical Association* 25: 397–406.

Frisch, R. 1926. On a problem in pure economics. In *Preferences, utility, and demand*, ed. J.S. Chipman, L. Hurwicz, M.K. Richter, and H.F. Sonnenschein, 1971. New York: Harcourt Brace Jovanovich.

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

Krantz, D.H., R.D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of measurement*. Vol. 3. New York: Academic Press.

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

Stevens, S.S. 1959. Measurement, psychophysics, and utility. In *Measurement. definitions and theories*, ed. C.W. Churchman and P. Ratoosh. New York: Wiley.

Von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*, 1956. Princeton: Princeton University Press.

Measurement Error Models

Han Hong

Abstract

Measurement error is important in econometric analysis. Its presence causes inconsistent parameter estimates. Under the classical measurement error assumption, instrumental variable methods can be used to eliminate the bias caused by measurement errors using a second measurement. This technique can be extended to polynomial regression models. For nonlinear models, deconvolution methods have been developed to cope with classical measurement errors. When the classical measurement error assumption is violated, auxiliary data-sets are usually needed to provide additional source of identification. When the true variable takes only discrete values, the mismeasurement problem takes the form of misclassification and requires special techniques.

Keywords

Attenuation bias; Auxiliary data; Deconvolution method; Generalized method of moments; Instrumental variables; Inverse probability weighting estimation; Linear models; Measurement error models; Misclassification; Nonlinear models; Permanent-income hypothesis; Polynomial regression; Semiparametric method; Sieve estimator

JEL Classifications

C10

Introduction

Many economic data-sets are contaminated by the mismeasured variables. Measurement error is one of the fundamental problems in empirical economics. The presence of measurement errors

causes biased and inconsistent parameter estimates, and leads to erroneous conclusions to various degrees in both linear and nonlinear econometric models. Techniques for addressing measurement error problems can be classified along two dimensions. Different techniques are used in linear models and in nonlinear models. Measurement error models that are valid under the classical measurement error assumption often are not applicable when the classical measurement error assumption does not hold.

Linear Models with Classical Measurement Errors

The classical measurement error assumption maintains that the measurement errors in any of the variables in a data-set are independent of all the true variables that are the objects of interest. The implication of this assumption in the linear least square regression model $y_i^* = x_i^{*\prime}$ is well understood and is usually described in standard econometrics textbooks. Under this assumption, measurement errors in the dependent variable $y_i = y_i^* + v_i$ do not lead to inconsistent estimates of the regression coefficients. Its only consequence is to inflate the standard errors of those regression coefficient estimates. On the other hand, independent errors that are present in the observations of the regressors $x_i = x_i^* + \eta_i$ lead to attenuation bias in simple univariate regression models and to inconsistent regression coefficient estimates in general. The importance of measurement errors in analysing the empirical implications of economic theories is highlighted in Milton Friedman's seminal book on the consumption theory of the permanent income hypothesis (Friedman 1957). In Friedman's model, both consumption and income consist of a permanent component and a transitory component that can arise from measurement errors or genuine fluctuations. The marginal propensity to consume relates the permanent component of consumption to the permanent income component. Friedman showed that, because of the attenuation bias, the slope coefficient of a regression of observed consumption on observed income would lead to an underestimate of the marginal propensity to consume.

Econometric work on linear models with classical independent additive measurement error dates back to Frish (1934), who derived bounds on the slope and the constant term. Instrumental variables (IV) is a popular method for obtaining consistent point estimators of the parameters of interest in this classical independent additive measurement error model. A valid instrument often comes from the second measurement of the error-prone true variable: $w_i = x_i^* + v_i$ which is subject to another independent measurement error v_i. The second measurement w_i is a valid instrument for the first measurement x_i because it is independent of both ε_i and η_i, but is correlated with the regressor x_i based on the first measurement.

The double-measurement instrumental variable method for linear regression models has been generalized by Hausman et al. (1991) to certain nonlinear regression models in which the regressors are polynomial functions of the error-prone variables. The following is a simplified version of the polynomial regression model that they considered:

$$y = \sum_{j=0}^{K} \beta_j z^j + r'\varphi + \varepsilon.$$

Among the two sets of regressors z and r, r is precisely observed but z is observed only with errors. In particular, two measurements of z, x and w, are observed which satisfy

$$x = z + \eta \text{ and } w = z + v.$$

An i.i.d. sample of observations is assumed to be available. Therefore we focus on identification of population moments. For convenience, assume that ε, η and v are mutually independent and they are independent of all the true regressors in the model.

First assume that $\varphi = 0$, then identification of β depends on population moments $\xi_j \equiv E(yz^j)$, $j = 0, \ldots, K$ and $\zeta_m \equiv Ez^m$, $m = 0, \ldots, 2K$, which are the elements of the population normal equations for solving for β. Except for ξ_0 and ζ_0, these moments depend on z which is not observed, but they can be

solved from the moments of observable variables Exw^{j-1}, Ew^l for $j = 0; \ldots; 2K$ and $Eyw^j, j = 0, \ldots, K$. Define $v_k = Ev^k$. Then the observable moments satisfy the following relations:

$$\begin{aligned} Exw^j &= E(z+\eta)(z+v)^j \\ &= E\sum_{l=0}^{j}\binom{j}{l}(z+\eta)z^l v^{j-1} \\ &= \sum_{l=0}^{j}\binom{j}{l}\zeta_{l+1}v_{j-l}, j \\ &= 1, 2K-1, \end{aligned} \tag{1}$$

and

$$Ew^j = E(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}z^l v^{j-1}$$

$$= \sum_{l=0}^{j}\binom{j}{l}\zeta_l v_{j-l}, j = 1, \ldots, 2K, \tag{2}$$

and

$$Eyw^j = Ey(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}yz^l v^{j-1}$$

$$= \sum_{l=0}^{j}\binom{j}{l}\xi_l v_{j-l}, j = 1, \ldots, K. \tag{3}$$

Since $v_1 = 0$, we have a total of $(5K - 1)$ unknowns in $\zeta_1, \ldots; \zeta_{2K}, \xi_1, \ldots, \xi_K$ and $v_2, \ldots, v_{2K}$. Equations (1), (3) and (4) give a total of $5K - 1$ equations that can be used to solve for these $5K - 1$ unknowns. In particular, the $4K - 1$ Eqs. in (1) and (3) jointly solve for $\zeta_1, \ldots; \zeta_{2K}, v_2, \ldots, v_{2K}$. Subsequently, given knowledge of these ζ's and v's, ξ's can then be recovered from Eq. (4). Finally, we can use these identified quantities of $\xi_j, j = 0, \ldots, K$ and $\zeta_m, m = 0, \ldots, 2K$ to recover the parameters β from the normal equations

$$\xi_l = \sum_{j=0}^{K}\beta_j\zeta_{j+l}, l = 0, \ldots, K.$$

When $\boldsymbol{\phi} \neq 0$, Hausman et al. (1991) noted that the normal equations for the identification of β and $\boldsymbol{\phi}$ depends on a second set of moments Eyr, Err^j and Erz^j, $j = 0, \ldots, K$, in addition to the first set of moments ξ's and ζ's. Since Eyr and Err^j can be directly observed from the data, it only remains to identify Erz^j, $j = 0, \ldots, K$. But these can be solved from the following system of equations, for $j = 0, \ldots, K$:

$$Erw^j = Er(z+v)^j = E\sum_{l=0}^{j}\binom{j}{l}rz^l v^{j-l}$$

$$= \sum_{l=0}^{j}\binom{j}{l}(Erz^l), \quad j = 0, \ldots, K.$$

In particular, using the previously determined v coefficients, the jth row of the previous equation can be solved recursively to obtain

$$Erz^j = Erw^j - \sum_{l=0}^{j-1}\binom{j}{l}(Ez^l l)v_{j-l}.$$

Once all these elements of the normal equations are identified, the coefficients β and φ can then be solved from the normal equations $[EyZ', Eyr]' = D½\beta', \boldsymbol{\phi}']$, where $Z = (1, z, \ldots, z^K)$ and $D = E[Z'r'), (Z'r')]$.

Nonlinear Model with Classical Measurement Errors

The deconvolution method is a useful technique to analyse general nonlinear model

$$Em(y^*; \beta) = 0$$

under the classical measurement error assumption with double measurements. These techniques are developed by Schennach (2004), Li (2002) and Taupin (2001). Suppose one knows the characteristic function $\psi_\eta(t) = Ee^{it\eta_i}$ of the errors η_i where only $y_i = y_i^* + \eta_i$ is observed and y^i A R^k. Then the characteristic function of y_i^* can be recovered from the ratio of the characteristic functions $\hat{\varphi}_y(t)$ and $\boldsymbol{\phi}_n(t)$ of y_i and η_i:

M

$$\hat{\varphi}_{y*}(t) = \hat{\varphi}_y(t)/\varphi_\eta(t).$$

where $\hat{\varphi}_y(t)$ can be estimated using a smooth version of $\frac{1}{n}\sum_{i=1}^{n} e^{ityi}$. Once the characteristic function of y^* is known, its density can be recovered from the inverse Fourier transformations

$$\hat{f}(y^*) = \left(\frac{1}{2\pi}\right)^k \int \hat{\varphi}_{\mathbf{y}*}(\mathbf{t}) e^{-iy*'\mathbf{t}} d\mathbf{t}.$$

For each β, a sample analog of the moment condition can then be estimated by

$$\int m(y^*;\beta)\hat{f}(y^*)dy^*.$$

A semiparametric generalized method of moment (GMM) estimator can be formed by minimizing over β a quadratic distance of the above estimated moment condition from zeros. Often, the characteristic function of the measurement errors $\boldsymbol{\phi}_n(t)$ might not be known. However, if two independent measurements of the latent true variable y^* with additive errors are observed and the errors are i.i.d, an estimate of $\hat{\varphi}_y(t)$ can be obtained using the two independent measurements.

For certain parametric families of the measurement error distribution, $\boldsymbol{\phi}(t)$ can be parameterized and its parameters can be estimated jointly with β. Hong and Tamer (2003) assume that the marginal distributions of the measurement errors are Laplace (double exponential) with zero means and unknown variances, and the measurement errors are independent of the latent variables and are independent of each other. Under these assumptions, they derive simple revised moment conditions in terms of the observed variables that lead to a simple estimator for nonlinear method of moment models with measurement error of the classical type when no additional data are available.

When the distributions of η are independent double Laplace, its characteristic function takes the form of

$$\varphi_\eta(t) = \prod_{j=1}^{k}\left(1+\frac{1}{2}\sigma_j^2 t_j^2\right)^{-1}.$$

Using this characteristic function, Hong and Tamer (2003) (Theorem 1) show that the moment condition $Em(y^*;\beta)$ can be translated into observable variable y as

$$Em(\mathbf{y}^*;\beta) = Em(\mathbf{y};\beta) + \sum_{l=1}^{k}\left(-\frac{1}{2}\right)^l \sum_{j_1<\cdots<jl}\cdots\sum \sigma_{j1}^2\cdots\sigma_{jl}^2 E\frac{\partial^{2l}}{\partial y_{j1}^2\cdots\partial y_{jl}^2} m(\mathbf{y};\beta).$$

For each candidate parameter value β, the right-hand side of the above can be estimated from the sample analog by replacing the expectation with the empirical sum. It can then be used to form a quadratic GMM objective function which can be used to estimate jointly β and the variance parameters $\sigma_j's$ of the double exponential distributions.

Non-classical Measurement Errors

The recent applied economics literature has raised concerns about the validity of the classical measurement error assumption. For example, in economic data it is often the case that data-sets rely on individual respondents to provide information. It may be hard to tell whether or not respondents are making up their answers and, more crucially, whether the measurement error is correlated with some of the variables. Studies by Bound and Krueger (1991), Bound et al. (1994) and Bollinger (1998) have all documented evidences of non-classical measurement errors. In order to obtain consistent estimates of the parameters β in the moment conditions $m(y^*;\beta)$, Chen et al. (2004, 2005) make use of an auxiliary data-set to recover the correlation between the measurement errors and the underlying true variables by estimating the conditional distribution of the measurement errors given the observed reported variables or proxy variables. In their model, the auxiliary data-set is a subset of the primary data, indicated by a dummy variable $D = 0$, which contains both the reported variable Y and the validated true variable Y^*. Y^* is not observed in the rest of the primary data-set ($D = 1$) which is not validated. The authors assume that the conditional distribution of the true variables given the reported variables can be recovered from the auxiliary data-set:

Assumption 1 $Y^* \perp D|Y$.

Under this assumption, an application of the law of iterated expectations gives

$$E[m(Y^*;\beta)] = \int g(Y;\beta)f(Y)dY \text{ where } g(Y;\beta) \times |Y, D = 0].$$

This suggests a semiparametric GMM estimator for the parameter β. For each value of β in the parameter space, the conditional expectation function $g(Y;\beta)$ can be nonparametrically estimated using the auxiliary data-set where $D = 0$.

Chen et al. (2005) suggest using sieve methods to implement this nonparametric regression. Let n denote the size of the entire primary data-set and let n_a denote the size of the auxiliary data-set where $D = 0$. Let $\{q_l(Y), l = 1, 2, .. .\}$ denote a sequence of known basis functions that can approximate any square-measurable function of X arbitrarily well. Also let

$$q^{k(n_a)}(Y) = \left(q_1(Y), \ldots, q_{k(n_a)}(Y)\right)' \text{ and}$$
$$Q_a = \left(q^{k(n_a)}(Y_{a1}), \ldots, q_{k(n_a)}(Y_{an_a})\right)'$$

for some integer $k(n_a)$, with $k(n_a) \to \infty$ and $k(n_a)/n \to 0$ when $n \to \infty$. In the above Y_{aj} denotes the jth observation in the auxiliary sample. Then for each given β, the first step nonparametric estimation can be defined as,

$$\hat{g}(Y;\beta) = \sum_{j=1}^{n_a} m\left(Y^*_{aj};\beta\right) q^{k(n_a)}\left(Y_{aj}\right) \times \left(Q_a'Q_a\right)^{-1} q^{k(n_a)}(Y).$$

A GMM estimator for β_0 can then be defined using a positive definite weighting matrix $\hat{W}$ as

$$\hat{\beta} = \arg\min_{\beta \in B} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}(Y_i;\beta)\right)' \hat{W} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}(Y_i;\beta)\right).$$

Chen et al. (2004) show that a proper choice of $\hat{W}$ achieves the semiparametric efficiency bound for the estimation of β. They called this estimator the 'conditional expectation projection estimator'.

Assumption (1) allows the auxiliary data-set to be collected using a *stratified sampling* design where a *non-random response-based subsample* of the primary data is validated. In a typical example of this stratified sampling design, we first oversample a certain subpopulation of the mismeasured variables Y, and then validate the true variables Y^* corresponding to this nonrandom stratified subsample of Y. It is very natural and sensible to oversample a sub-population of the primary data-set where more severe measurement error is suspected to be present. Assumption 3.1 is valid as long as, in this sampling procedure of the auxiliary data-set, the sampling scheme of Y in the auxiliary data is based only on the information available in the distribution of the primary data-set $\{Y\}$. For example, one can choose a subset of the primary data-set $\{Y\}$ and validate the corresponding $\{Y^*\}$, in which case the Y's in the auxiliary data set are a subset of the primary data Y. The stratified sampling procedure can be illustrated as follows. Let U_{pi} be i.i.d $U(0,1)$ random variables independent of both Y_{pi} and Y^*_{pi}, and let $T(Y_{pi}) \in (0,1)$ be a measurable function of the primary data. The stratified sample is obtained by validating every observation for which $U_{pi} < T(Y_{pi})$. In other words, $T(Y_{pi})$ specifies the probability of validating an observation after Y_{pi} is observed.

A special case of assumption 3.1 is when the auxiliary data is generated from the same population as the primary data, where a full independence assumption is satisfied:

Assumption 2 $Y, Y^* \perp D$.

This case is often referred to as a validation sample. Semiparametric estimators that make use of a validation sample include Carroll and Wand (1991), Sepanski and Carroll (1993), Lee and Sepanski (1995), and the recent work of Devereux and Tripathi (2005). Interestingly, in the case of a validation sample, Lee and Sepanski (1995) suggest that the nonparametric estimation of the conditional expectation function $g(Y;\beta)$ can be replaced by a finite dimensional linear projection $h(Y;\beta)$ into a fixed set of functions of Y. In other words, instead of requiring that $k(n_a) \to \infty$ and $k(n_a)=n \to 0$, we can hold $k(n_a)$ to be a fixed constant in the above least square regression for $\hat{g}(Y;\beta)$. Lee and Sepanski (1995) show that this

will still produce a consistent and asymptotically normal estimator for β as long as the auxiliary sample is also a validation sample that satisfies assumption 2. However, if the auxiliary sample satisfies assumption 1 but not assumption 2, then it is necessary to require $k(n_a) \to \infty$ to obtain consistency. Furthermore, even in the case of a validation sample, requiring $k(n_a) \to \infty$ typically results in a more efficient estimator for β than a constant $k(n_a)$.

An alternative consistent estimator that is valid under assumption 1 is based on the *inverse probability weighting* principle which provides an equivalent representation of the moment condition $Em(y^*;\beta)$. Define $p(Y) = p(D = 1|Y)$,

$$Em(y;\beta) = E\left[m(Y^*;\beta_0)\frac{1-p}{1-p(Y)}|\, D=0\right].$$

To see this, note that

$$\begin{aligned} E\left[m(Y^*;\beta_0)\frac{1-p}{1-p(Y)}|\, D=0\right] &= \int m(Y^*;\beta_0)\frac{1-p}{1-p(Y)}\frac{f(Y)(1-p(Y))f(Y^*|\, Y,D=0)}{1-p}dY^*dY \\ &= \int m(Y^*;\beta_0)f(Y^*|\, Y)f(Y)dY^*dY = Em(y^*;\beta), \end{aligned}$$

where the third equality follows from assumption 3.1 that $f(Y^*|Y, D = 0) = f(Y^*|Y)$.

This equivalent reformulation of the moment condition $Em(Y^*;\beta)$ suggests a two-step inverse probability weighting estimation procedure. In the first step, one typically obtains a parametric or nonparametric estimate of the so-called propensity score $\hat{p}(Y)$ using, for example, a logistic binary choice model with a flexible functional form. In the second step, a sample analog of the re-weighted moment conditions is computed using the auxiliary data-set:

$$\hat{g}(\beta) = \frac{1}{n_a}\sum_{j=1}^{n_a} m\left(Y_j^*;\beta\right)\frac{1}{1-\hat{p}\left(Y_j\right)}.$$

This is then used to form a quadratic norm to provide a GMM estimator:

$$\hat{\beta} = \arg\min_{\beta}\hat{g}(\beta)\, W_n\hat{g}(\beta).$$

Interestingly, an analog of the conditional independence assumption 1 is also rooted in the program evaluation literature and is typically referred to as the assumption of un-confoundedness, or selection based on observables. Semiparametric efficiency results for the mean treatment effect parameters to nonlinear GMM models have been developed by, among other, Robins et al. (1992), Hahn (1998), Hirano et al. (2003) and Imbens et al. (2005). Many of the results presented here generalize these results for the mean treatment effect parameters to nonlinear GMM models.

Misclassification of Binary of Discrete Variables

Measurement problems on binary or discrete variables usually take the form of *mis-classification*: for example, a unionized worker might be mis-classified as one who is not unionized. When the variable of interest and its measurement are both binary, the measurement error can not be independent of the true binary variable. Typically, mis-classification introduces a negative correlation, or mean reversion, between the errors and the true values. Estimation methods that address the mis-classification problem have been developed by, among others, Abrevaya et al. (1998), Manski and Horowitz (1995), Molinari (2005) and Mahajan (2006).

In particular, the recent work by Mahajan (2006) studies a nonparametric regression model where one of the true regressors is a binary variable:

$$y = g(x^*, z) + \varepsilon \quad \text{where } E(\varepsilon|\, x^*, z) = 0.$$

Instead of observing x^*, the researchers are able only to observe a potentially misreported

binary value x of x^*. In the rest of this section we present the identification and estimation results developed in Mahajan (2006).

Mahajan (2006) assumes that, in addition, another random variable v is observed such that the following four assumptions hold.

Assumption 3 $E(y|\, x^*, z, x, v) = g(x^*, z)$.

Assumption (3) requires that conditional on the true variable x^*, the measurement error $x - x^*$ does not provide additional information about the outcome variable y. It also requires that v satisfies the following additional assumptions.

Assumption 4 $x \perp v|x^*\,, z,$

and for $\eta_2^*(z,v) = P(x^* = 1|\, z, v)$,

Assumption 5 $\eta_2^*(z,v)$ is a non-trivial function of v.

Mahajan (2006) calls the variable v an *instrument like variable* that is conditionally independence of the outcome y (assumption 3) and of the misreported value x (assumption 4), but is correlated with x^* given z (assumption 5). Assumption 3 is similar to the exclusion restriction for instrument variables in standard linear models. Assumption 5 is analogous to the requirement that an instrument should be correlated with regressors. Because of assumption 4, assumption 5 implies that $\eta_2(z, v) = P(x = 1|\, z, v)$ is also a non-trivial function of v given z.

In addition, Mahajan (2006) also imposes the following monotonicity assumption to restrict the extent of misclassification:

Assumption 6 Define $\eta_0(z) = P(x = 1|x^* = 0, z)$, and $\eta_1(z) = P(x = 0|x^* = 1, z)$. $\eta_0(z) + \eta_1(z) < 1$.

This assumption is innocuous since it can almost certainly be satisfied by relabelling the binary variables. Under these assumptions, Mahajan (2006) demonstrates that the regression function $g(x^*, z)$ can be nonparametrically identified. To see this, note that $\eta^2(z,v)$ is observable and note the following relations:

$$
\begin{aligned}
&E(x|\, z, v) \equiv \eta_2(z, v) = (1 - \eta_1(z))\eta_2^*(z, v) \\
&+\eta_0(z)\big(1 - \eta_2^*(z, v)\big)E(x|\, z, v) \\
&= g(1, z)\eta_2^*(z, v) + g(0, z)\big(1 - \eta_2^*(z, v)\big) \\
&\times E(yx|\, z, v) = g(1, z)(1 - \eta_1(z))\eta_2^*(z, v) \\
&+g(0, z)\eta_0(z)\big(1 - \eta_2^*(z, v)\big)
\end{aligned}
$$

Suppose v takes n_v values. For each z, $\eta_0(z)$, $\eta_1(z)$, $g(0, z)$, $g(1, z)$ and $\eta_2^*(z, v)$ are unknown. There are $4 + n_v$ parameters. There are $3n_v$ equations. Therefore, as long as $n_v \geq 2$, all the parameters can possibly be identified. Intuitively, if $\eta_2^*(z, v)$ is known, the second moment condition $E(y|z,v)$ identifies $g(1,z)$ and $g(0,z)$. Information from the other moment conditions also allows one to identify both $\eta_1(z)$ and $\eta_0(z)$.

A constructive proof is given in Mahajan (2006) using the above three moment conditions. First of all, using the first moment condition

$$
\eta_2^*(z, v) = \frac{\eta_2(z, v) - \eta_0(z)}{1 - \eta_0(z) - \eta_1(z)}.
$$

If this is substituted into the next two moment conditions, then one can write

$$
\begin{aligned}
&E(y|\, z, v) = g(0, z) + (g(1, z) - g(0, z)) \\
&\times \frac{\eta_2(z, v) - \eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} = g(0, z) \\
&- \frac{(g(1, z) - g(0, z))\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} \\
&+ \frac{g(1, z) - g(0, z)}{1 - \eta_0(z) - \eta_1(z)}\eta_2(z, v) \\
&E(yx|\, z, v) = g(0, z)\eta_0(z) \\
&-[g(1, z)1 - \eta_1(z) - g(0, z)\eta_0(z)] \\
&\times \frac{\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} \\
&+ \frac{[g(1, z)(1 - \eta_1(z)) - g(0, z)\eta_0(z)]}{1 - \eta_0(z) - \eta_1(z)}\eta_2(z, v) \\
&= -\frac{(g(1, z) - g(1, z))\eta_0(z)(1 - \eta_1(z))}{1 - \eta_0(z) - \eta_1(z)} \\
&+ \frac{[g(1, z)(1 - \eta_1(z)) - g(0, z)\eta_0(z)]}{1 - \eta_0(z) - \eta_1(z)}\eta_2(z, v)
\end{aligned}
$$

Mahajan (2006) suggests that, if one runs a regression of $E(y|z, v)$ on $\eta_2(z, v)$ and runs a regression of $E(yx|z, v)$ on $\eta_2(z, v)$, then one can recover the intercepts and the slope coefficients:

$$\begin{aligned} a &= g(0,z) - \frac{(g(1,z) - g(0,z))\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} b \\ &= \frac{g(1,z) - g(0,z)}{1 - \eta_0(z) - \eta_1(z)} c = g(0,z)\eta_0(z) \\ &-[g(1,z)(1 - \eta_1(z)) - m(0,z)\eta_0(z)] \\ &\times \frac{\eta_0(z)}{1 - \eta_0(z) - \eta_1(z)} = \\ &- \frac{(g(1,z) - g(1,z))\eta_0(z)(1 - \eta_1(z))}{1 - \eta_0(z) - \eta_1(z)} d \\ &= \frac{[g(1,z)(1 - \eta_1(z)) - g(0,z)\eta_0(z)]}{1 - \eta_0(z) - \eta_1(z)}. \end{aligned}$$

Therefore, one can write

$$a = m(0,z) - \eta_0(z)b \tag{4}$$

and

$$c = m(0,z)\eta_0(z) - d\eta_0(z) \tag{5}$$

$$c = -b(1 - \eta_1(z))\eta_0(z). \tag{6}$$

Equation (4) can be used to concentrate out $m(0, z)$. One can then substitute it into (5) and make use of (6) to write

$$\begin{aligned} &(a + \eta_0(z)b)\eta_0(z) - d\eta_0(z) \\ &= -d(1 - \eta_1(z))\eta_0(z). \end{aligned}$$

Then we can factor out $\eta_0(z)$ and rearrange:

$$1 - \eta_1(z) + \eta_0(z) = \frac{d - a}{b}. \tag{7}$$

Now we have two Eqs. (6) and (7) in two unknowns 1 - $\eta_1(z)$ and $\eta_0(z)$. Obviously the solutions to this quadratic system of equation is unique only up to an exchange between 1 - $\eta_1(z)$ and $\eta_0(z)$. However, assumption 6 rules out one of these two possibilities and allows for point identification. Hence Mahajan (2006) demonstrates that the model is identified.

Mahajan (2006) further develops his identification strategy into a nonparametric estimator, and also develops a semiparametric estimator for a single index model.

Conclusion

Despite numerous articles that have been written on the topic of measurement errors in econometrics and statistics over the years, there are still many unresolved important qsts that are related to models of measurement errors. For example, the implications of measurement errors and data contaminations on complex structural models in labour economics and industrial organization are yet to be understood and studied. Recent empirical studies of precautionary saving and the permanent income hypothesis make use of panel data to address the issue of measurement errors (see, for example, Parker and Preston 2005). Also, it is often the case that not all variables are validated in auxiliary data-sets. How to make use of partial information in validation studies is also an open qst.

See Also

- ▶ Econometrics
- ▶ Efficiency Bounds
- ▶ Linear Models
- ▶ Semiparametric Estimation

Acknowledgments The author acknowledges generous research support from the NSF (SES-0452143) and the Sloan Foundation.

Bibliography

Abrevaya, J., J. Hausman, and F. Scott-Morton. 1998. Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics* 87: 239–269.

Bollinger, C. 1998. Measurement error in the current population survey: A nonparametric look. *Journal of Labor Economics* 16: 576–594.
Bound, J., C. Brown, G. Duncan, and W. Rodgers. 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics* 12: 345–368.
Bound, J., and A. Krueger. 1991. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right. *Journal of Labor Economics* 12: 1–24.
Carroll, R., and M. Wand. 1991. Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society* 53: 573–585.
Chen, X., H. Hong, and E. Tamer. 2005. Measurement error models with auxiliary data. *Review of Economic Studies* 72: 343–366.
Chen, X., Hong, H. and Tarozzi, A. 2004. Semiparametric efficiency in GMM models nonclassical measurement errors. Working paper, Duke University and New York University.
Devereux, P. and Tripathi, G. 2005. Combining datasets to overcome selection caused by censoring and truncation in moment bases models. Working paper, University of Connecticut and UCLA.
Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.
Frish, R. 1934. *Statistical confluence study.* Oslo: University Institute of Economics.
Hahn, J. 1998. On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66: 315–332.
Hausman, J., H. Ichimura, W. Newey, and J. Powell. 1991. Measurement errors in polynomial regression models. *Journal of Econometrics* 50: 271–295.
Hirano, K., G. Imbens, and G. Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71: 1161–1189.
Hong, H., and E. Tamer. 2003. A simple estimator for nonlinear error in variable models. *Journal of Econometrics* 117: 1–19.
Imbens, G., Newey, W. and Ridder, G. 2005. Mean-squared-error calculations for average treatment effects. Working paper, Harvard University, MIT and USC.
Lee, L., and J. Sepanski. 1995. Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association* 90(429): 130–140.
Li, T. 2002. Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics* 110: 1–26.
Mahajan, A. 2006. Identification and estimation of regression models with misclassification. *Econometrica* 74: 631–665.
Manski, C., and J. Horowitz. 1995. Identification and robustness with contaminated and corrupted data. *Econometrica* 63: 281–302.
Molinari, F. 2005. Partial identification of probability distributions with misclassified data. Working paper, Cornell University.
Parker, J., and B. Preston. 2005. Precautionary savings and consumption fluctuations. *American Economic Review* 95: 1119–1144.
Robins, J., S. Mark, and W. Newey. 1992. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48: 479–495.
Schennach, S. 2004. Estimation of nonlinear models with measurement error. *Econometrica* 72: 33–75.
Sepanski, J., and R. Carroll. 1993. Semiparametric quasi-likelihood and variance estimation in measurement error models. *Journal of Econometrics* 58: 223–256.
Taupin, M. 2001. Semiparametric estimation in the nonlinear structural errors-in-variables model. *Annals of Statistics* 29: 66–93.

Measurement of Economic Growth

C. H. Feinstein

Economic growth is most commonly defined in terms of the rate of change in some measure of national product per head of population at constant prices. It is thus appropriate to consider four issues arising from this definition: the evolution of attempts to measure economic growth; the conceptual basis of the standard definition; various objections to this, and possible alternative concepts; and the actual historical record of selected nations with respect to their rate of economic growth.

The earliest estimates of national income were made at the end of the 17th century by Petty, King and Davenant for England, and by Boisguillebert and Vauban for France. Their main concern was the tax revenue which might be collected from this income, and their estimates were primarily designed to provide a static picture of its level and distribution at a point in time. However, even at this initial stage Gregory King had the inspiration and ability to undertake a pioneering exercise in international comparison: he estimated the national income of England, France and Holland in 1688 and 1697, and was thus able to

assess the effects of the war on the growth of each of the main belligerents.

Little further progress was made until the early 19th century. Then Joseph Lowe, building on a recent study by Patrick Colquhoun, constructed estimates of the national income of Great Britain for four dates spanning the Napoleonic wars and their immediate aftermath. His crucial innovation was to make his calculation in 'money of uniform value', thereby enabling his readers to see the increase in incomes 'without the perplexity attendant on a difference in the value of our currency' (Lowe 1823, p. 36). Later in the century further impetus was given to the study of economic growth, most notably by the statistical enquiries of Tucker in the United States, Giffen in the United Kingdom and Coghlan in Australia.

In the present century major advances were made in the clarification of the underlying theoretical concepts and in the empirical definition and measurement of real incomes. The outstanding figures include Pigou, Hicks, Bowley, Clark and Stone in the United Kingdom; and King, Kuznets, Gilbert and Jaszi in the United States. For a full account of these and earlier developments see Studenski (1958). By mid-century the adoption of Keynesian macroeconomic policies in industrialized market economies, the spread of planning in socialist economies and the concern with low incomes in numerous less-developed countries had combined to produce a worldwide appreciation of the advantages of national accounts as a framework for planning and policy-making. Across the globe, private estimates gave way to official series, and the recommendations of international organizations encouraged the adoption of comparable concepts and methods. Systematic measurement and analysis of economic growth had come of age.

The concept most widely used as the basis for such studies is gross domestic product at constant prices: *real GDP*. Viewed from the expenditure side this may be regarded as the aggregate expenditure by the residents of a country on final goods and services for private consumption, for investment at home and abroad, and for government expenditure on health, education, defence and other services. There are equivalent definitions in terms of aggregate factor incomes and aggregate output (value added). Economic growth would thus be measured by the rate of increase in real GDP, or if it is desired to adjust for the effect of changes in population, of *real GDP per head*. However, if the central concern is with the rate of growth of efficiency or productivity, the appropriate measure would be *real GDP per worker*, or some equivalent estimate of output per unit of input.

A number of alternative concepts may also be used. For example, gross *national* product (GNP), which adds to GDP the income obtained by nationals from labour or property ownership outside the country, and deducts the income of foreign nationals arising within the country. Or *net* domestic (national) product, which deducts an allowance for the value of capital assets consumed in the course of production of GDP (GNP). All of these may be valued either at market prices or at factor cost. These variants will have different levels, but their rates of growth will not normally differ significantly. The two aspects which in practice are thus of critical importance for the measurement of economic growth are the conceptual basis of the GDP series and the correction for changes in prices.

The former raises many complex issues which cannot be examined here, but there are a few points which merit particular attention in the context of any consideration of economic growth. (For more detailed discussion of these issues see Moss 1973, and Usher 1980.) The most important is that calculations based on the change in GDP are specifically designed to measure only the growth in certain aspects of economic welfare, or – on an alternative interpretation – in the productive capacity of the economy. Changes in such non-economic aspects of the quality of life as social justice, the pleasure of watching a golden sunset or the personal benefits of living longer are completely omitted. GDP is thus, at best, a measure of some important aspects of economic welfare – it is in no sense a comprehensive measure of 'well-being'. Furthermore, even within its chosen field of economic welfare the standard measures of GDP suffer from a number of major deficiencies.

First, there are certain costs directly associated with economic growth, such as pollution of the

environment, loss of leisure, and longer journeys to work. All of these costs may rise as output expands, but they are not entered as negative items in the compilation of GDP, and many contend that the apparent gains from the growth thus recorded are wholly or largely illusory. Secondly, there are some parts of marketed output which enter into GDP but which are not generally regarded as desirable in themselves: the services and equipment of the armed forces and police are the most commonly cited examples of such 'regrettable necessities'. The merits of this procedure have frequently been challenged on the grounds that an increase in, say, defence spending should not be treated as a contribution to economic welfare.

Thirdly, it is general practice to cover only those goods and services which are traded in the market. A few exceptions are made, notably for agricultural output which is consumed by its producers – an important feature in many subsistence economies. But one much-discussed item which is not normally included is the substantial amount of unpaid work done within the household. Measured growth may thus increase simply because households start to purchase certain goods and services which they had previously provided for themselves. Equally, it may decline as the reverse process occurs; for example, as domestic servants are replaced by members of the household working with the aid of a variety of consumer durables. Finally, it must always be recognized that changes in total (or average) GDP take no account of changes in the distribution of the goods and services between different individuals or groups within the community. An overall increase in real GDP may be accompanied by an absolute decline in the incomes of the lowest income group, and its benefits appraised accordingly.

A number of attempts have been made to construct alternative measures which allow for one or more of these objections to GDP. The best known is probably the 'measure of economic welfare' compiled by Nordhaus and Tobin (1972). This made a number of adjustments to the United States estimates of real GNP for 1929–65, including an addition for the value of leisure and of non-market work, and a deduction for expenditure on defence and other 'regrettable necessities' and for the costs of urbanization.

However, none of these variants has as yet won general recognition or been published on a regular basis. This is mainly because they pose enormous practical problems for the statisticians. How, for example, should leisure-time be valued, or the unpaid work of a housewife? If defence should be excluded as a regrettable necessity, should the services of doctors and dentists be similarly excluded because they are required only to provide relief from pain or illness? In the absence of agreement on how such difficulties might best be resolved it has been found helpful to concentrate attention on the narrower but less ambiguous area based predominantly on marketed activity. Despite its acknowledged limitations, GDP provides a reasonably consistent and reliable measure of a substantial component of economic welfare and thus remains the standard basis for comparisons of economic growth over time or between nations.

With regard to the second aspect of real GDP noted above – the deflation for changing prices – there are again major conceptual issues. The first of these arises from what is known as the 'index number problem'. Its essence is that the collection of prices (per ton of steel, per metre of cloth, etc.) which must be used for the price deflation must be weighted according to the pattern of output or expenditure in some particular 'base' year. The result given by a base year at the beginning of the period over which growth is to be measured may differ markedly from that obtained when the chosen year is at the end of the period. It can be shown that the former will typically give a higher rate of growth of real GDP than the latter; and the problem is a disturbing one because it is normally the case that the discrepancy is greatest in precisely those episodes of rapid growth and structural change which give the study of economic growth some of its most fascinating subjects (e.g. the industrial revolution in England or the period of the first two Five- Year Plans in the USSR). There is thus an inescapable element of ambiguity at the heart of the measurement of economic growth.

The second problem is created by the fact that the composition of the goods and services covered

Measurement of Economic Growth, Table 1 Growth of real GDP per head of population, 1820–1979 (Annual average compound growth rates) (Source Maddison (1982, p. 44))

	1820–70	1870–1913	1913–50	1950–73	1973–79
Australia	..	0.6	0.7	2.5	1.3
Austria	0.7	1.5	0.2	5.0	3.1
Belgium	1.9	1.0	0.7	3.6	2.1
Canada	..	2.0	1.3	3.0	2.1
Denmark	0.9	1.6	1.5	3.3	1.8
France	1.0	1.5	1.0	4.1	2.6
Germany	1.1	1.6	0.7	5.0	2.6
Italy	..	0.8	0.7	4.8	2.0
Japan	0.0	1.5	0.5	8.4	3.0
Netherlands	1.5	0.9	1.1	3.5	1.7
Norway	1.0	1.3	2.1	3.1	3.9
Sweden	0.6	2.1	2.2	3.1	1.5
United Kingdom	1.5	1.0	0.9	2.5	1.3
United States	1.4	2.0	1.6	2.2	1.9

by GDP changes over time, and the longer the period, the greater the changes are likely to be. Many items purchased at the beginning will no longer be produced or bought at the end, and those which have taken their place will not have been available at the beginning. The accuracy of any correction for price changes will inevitably be affected by this characteristic of economic growth. Even where an item is produced over the entire period, there are likely to be improvements in quality which cannot easily be separated from the associated changes in prices and quantity; and it is often suggested that the rise in prices may be overstated, and the growth of real GDP correspondingly understated, by failure to identify what are actually changes in quality (Table 1).

Notwithstanding all these and other reservations, the measurement of economic growth has continued to attract widespread interest. We therefore conclude this brief review of the topic by referring the reader to the table setting out the long-term historical record for a number of the leading developed countries.

See Also

- ▶ Growth Accounting
- ▶ Index Numbers
- ▶ International Income Comparisons
- ▶ National Income
- ▶ Social Accounting

Bibliography

Lowe, A. 1823. *The present state of England in regard to agriculture, trade and finance*. London.
Maddison, A. 1982. *Phases of capitalist development*. Oxford: Oxford University Press.
Moss, M. (ed.). 1973. *The measurement of economic and social performance*, Studies in Income and Wealth, vol. 38. New York: NBER.
Nordhaus, W.D., and J. Tobin. 1972. Is growth obsolete? In *Economic growth*, Fiftieth Anniversary Colloquium. New York: NBER.
Studenski, P. 1958. *The income of nations*. New York: New York University Press.
Usher, D. 1980. *The measurement of economic growth*. Oxford: Basil Blackwell.

Measurement, Theory of

R. Duncan Luce and Louis Narens

Abstract

Physical measurement, as embodied in dimensional analysis, consists of interlocked, qualitative, ordered structures. Analogous

approaches to behavioural science are outlined for sensory scaling such as loudness, utility of uncertain alternatives, and qualitative foundations of probability. In many cases, they are an ordered structure with a binary operation for combining elements and a Cartesian product where each factor affects the ordering of the attribute. Axioms sufficient for measurement – for the existence of a homomorphism onto the positive real numbers – are mentioned, and their uniqueness (scale type – for example, ratio, interval, ordinal) is formulated qualitatively in terms of the structure's symmetries (or automorphisms).

Keywords

Additivity; Allais Paradox; Averaging; Completeness; Conditional probability; De Finetti, B.; Extended sure-thing principle; Independence; Invariance; Kolmogorov, A. N.; Mass measurement; Measurement, theory of; Monotonicity; Preference reversals; Probability; Ratio scale; Representation; Scale of measurement; Subjective expected utility; Subjective probability; Transitivity; Unboundedness; Unconditional probability

JEL Classifications

C0

Most mathematical sciences rest upon quantitative models, and the theory of measurement is devoted to making explicit the qualitative assumptions that underlie them. This is accomplished by first stating the qualitative assumptions – empirical laws of the most elementary sort – in axiomatic form and then showing that there are structure preserving mappings, often but not always isomorphisms, from the qualitative structure into a quantitative one. The set of such mappings forms what is called a 'scale of measurement'.

A theory of the possible numerical scales plays an important role throughout measurement – and therefore throughout science. Just as the qualitative assumptions of a class of structures narrowly determine the nature of the possible scales, so also the nature of the underlying scales greatly limits the possible qualitative structures that give rise to such scales. Two major themes of this entry reflect research results of the 1970s and 1980s: (*a*) the possible scales that are useful in science are necessarily very limited; (*b*) once a type of scale is selected (or assumed to exist) for a qualitative structure, then a great deal is known about that structure and its quantitative models. A third theme concerns applications of these ideas to the behavioural sciences, especially to utility theory and psychophysics from 1980 onward.

There are several general references to the axiomatic theory. Perhaps the most elementary and the one with the most examples is Roberts (1979). Pfanzagl (1968) and Krantz et al. (1971) are on a par, with the latter more comprehensive. Narens (1985), which is the mathematically most sophisticated, covers much of the basic material mentioned here. Later additions are: Luce et al. (1990), which has much in common with Narens (1985); Suppes et al. (1989), which is focused on geometric representations and probability generalizations; Narens (2007), which is a more narrowly focused introductory book with examples mainly from psychophysics; and Suppes (2002). Mostly, we cite only references not included in wither Krantz et al. (1971) or Narens (1985).

Axiomatizability

The qualitative situation is usually conceptualized as a relational structure $X = \langle X, S_0, S_1, \ldots \rangle$, where the $S_0, S_1, \ldots$ are finitary relations on X. The number of relations can be either finite or infinite, but in applications almost always finite. X is called the domain of the structure and the S_i its primitive relations. In most applications, S_0 will be some type of ordering relation that is usually written as $\succsim$. The following are some examples of qualitative structures used in measurement situations.

The first goes back to Helmholtz (see sect. "Axiomatization of Concatenation Structures"). It has for its domain a set X of objects with the properties like those of mass. There are two primitive relations. The first, $\succsim$, is a binary ordering according to mass (which may be determined, for

example, by using an equal-arm pan balance so that $x \succsim y$ means that the pans either remain level or the one containing x drops). The second is a binary operation o, which formally is a ternary relation. For mass it is empirically defined as follows: if x and y are placed in the same pan and are exactly balanced by z, then we write $xoy \sim z$, where $\sim$ means equivalence in the attribute. Other interpretations of the primitives of $\langle X, \succsim, \text{o} \rangle$ can be found in the above references. Axiomatic treatments of the structure $\langle X, \succsim, \text{o} \rangle$ are discussed in sect. "Axiomatization of Concatenation Structures".

A second example is from economics. Suppose that $C_1, \ldots, C_n$ are sets each consisting of different amounts of a commodity, and $\succsim$ is a preference ordering exhibited by a person or an institution over the set of possible commodity bundles $C = \Pi_i C_i$. $\langle C, \succsim \rangle$ is called a *conjoint structure*, and axioms about it are given that among other things induce an ordering, $\succsim_i$, of an individual's preferences for the commodities associated with each component i.

A third example, due to *B*. de Finetti, has as its domain an algebra of subsets, called 'events', of some non-empty set Ω. The primitives of the structure consist of an ordering relation $\succsim$ of 'at least as likely as', the events Ω and $\varnothing$ and the set theoretical operations of union $\cup$, intersection $\cap$, and complementation $\neg$.

The relational structure

$$\mathcal{P} = \langle \mathcal{E}, \succsim, \Omega, \varnothing, \cup, \cap, \neg \rangle \qquad (1)$$

is intended to characterize qualitatively probability-like situations. The primitive $\succsim$ can arise from many different processes, depending upon the situation. In one, which is of considerable importance to Bayesian probability theorists and statisticians, $\succsim$ represents a person's ordering of events according to how likely they seem, using whatever basis he or she wishes in making the judgements.

In such a case, $\mathcal{P}$ is thought of as a *subjective or personal probability structure*. In another, $\succsim$ is an ordering of events based on some probability model for the situation (possibly one coupled with estimated relative frequencies), as in much of classical probability theory.

Ordered Structure

Weak Order, Dedekind Completeness, and Unboundedness

Two types of 'quantitative' representations have played a major role in science: systems of coordinate geometry and the real number system (the latter being the one-dimensional specialization of the former). Results about the former are in Suppes et al. (1989), but our focus here is the latter. The absolutely simplest case, included in all of the above examples, is the order-preserving representation φ of $\langle X, \succsim \rangle$ into $\langle \mathbb{R}, \geq \rangle$, where $\mathbb{R}$ denotes the real numbers. An immediate implication is that $\succsim$ must be transitive, reflexive, and connected (for all x and y, either $x \succsim y$ or $y \succsim x$). Such relations are given many different names including *weak order*. An antisymmetric weak order is called a total or *simple order*. There has been much empirical controversy about the transitivity of $\succsim$, with the most recent Bayesian analyses favouring transitivity of $\succ$ but not of $\sim$ (Myung et al. 2005). Some doubt has been expressed about completeness. Nevertheless, most of the well-developed measurement-theoretic techniques assume both the completeness and transitivity of $\succsim$ as idealizations.

G. Cantor showed that for $\langle X, \succsim \rangle$ to be so represented, necessary and sufficient conditions are that $\succsim$ be a weak order and that there be a finite or countable subset Y of X that is order dense in X (that is, for each $x \succ z$ there exists a y in Y such that $x \succsim y \succsim z$). For many purposes, this subset plays the same role as do the rational numbers within the system of real numbers.

In order for the representation to be onto either $\langle \mathbb{R}, \geq \rangle \langle \mathbb{R}^+, \geq \rangle$, where $\mathbb{R}^+$ denotes the positive real numbers, which often happens in physical measurement, two additional conditions are necessary and sufficient: *Dedekind completeness* (each non-empty bounded subset of X has a least upper bound in X) and *unboundedness* (there is neither a least nor a greatest element).

In measurement axiomatizations, one usually does not postulate a countable, order-dense subset, but derives it from axioms that are intuitively more natural. For example, with a binary operation of combining objects, order density follows

from a number of properties including an *Archimedean axiom* which states in some fashion that no object is either infinitely larger than or infinitesimally close to another object. When the structure is Dedekind complete and the operation is monotonic, it is also Archimedean. Dedekind completeness and Archimedeaness are what logicians call 'second-order axioms', and in principle they are incapable of direct empirical verification.

The most fruitful and intensively examined measurement structures are those with a weak ordering $\succsim$ and an associative, positive binary operation o that is strictly monotonic ($x \succsim y$ iff $xoz \succsim yoz$). They have been the basis of much physical measurement. However, for much of the 20th century they played little role in the behavioural and social sciences but, as seen in sects. "Interlocked Structures and Applications to Utility Theory" and "Other Applications of Behavioural Interest", since the 1997s such operations have come to be useful. The development of a general non-associative and non-positive ($x \succsim xoy$ for some x and y) theory began in 1976, and it is moderately well understood in certain situations having many symmetries. (Technically, symmetries or automorphisms of the structure are isomorphic transformations of the structure onto itself.) This, and its specialization to associative structures, is the focus of sect. "Scale Types".

Representations and Scales

A key concept in the theory of measurement is that of a *representation*, which is defined to be a structure preserving map φ of the qualitative, weakly ordered relational structure $\mathcal{X}$ into a quantitative one, $\mathcal{R}$, in which the domain is a subset of the real numbers. Representations are either isomorphisms or homomorphisms. The latter are used in cases where equivalences play an important role (for example, conjoint structures where trade-offs between components are the essence of the matter), in which case equivalence classes of equivalent elements are assigned the same number. We say φ is a *$\mathcal{R}$-representation for $\mathcal{X}$*.

From 1960 to 1990, measurement theorists were largely focused on certain types of qualitative structures for which numerical representations exist. The questions faced are two. The first, the 'existence' problem, is to establish that the set of $\mathcal{R}$-representations is non-empty for $\mathcal{X}$. Cantor's conditions above establish existence of a numerical representation of any weak order. The second, the 'uniqueness', problem is to describe compactly the set of all $\mathcal{R}$-representations. Several examples are cited. Since 1990, the focus has been increasingly on applying these insights to behaviour. We cite aspects of utility theory, global psychophysics, and probability.

For the qualitative mass structure $\mathcal{X} = \langle X, \succsim \mathrm{o} \rangle$ described previously, the qualitative representing structure is taken to be $\mathcal{R} = \langle \mathbb{R}^+, \geq, + \rangle$ where $\geq$ and + have their usual meanings in $\mathbb{R}^+$. The set of $\mathcal{R}$-representations of $\mathcal{X}$ consist of all functions φ from X into $\mathbb{R}^+$ such that for each x and y in X,

(i) $X \succsim y$ iff $\varphi(x) \geq \varphi(y)$, and
(ii) $\varphi(xoy) = \varphi(x) + \varphi(y)$.

Such a function is called a *homomorphism for $\mathcal{X}$*, and the set of all of them is called a *scale (for $\mathcal{X}$)*. In addition to Helmholtz, others – including Hölder, Suppes, Luce and Marley, and Falmagne – have stated axioms about the primitives that are sufficient to show the existence of such homomorphisms and to show the following uniqueness theorem: any two homomorphisms φ and ψ are related by positive multiplication, that is, there is some real $r > 0$ such that $\psi = r\varphi$. In the language introduced by Stevens (1946), such a form of measurement is said to form a 'ratio scale'. For cases where is an operation (defined for all pairs), Alimov (1950) and Roberts and Luce (1968) gave necessary and sufficient conditions for such a representation. Such a complete characterization as this one is rather unusual in measurement; sufficient conditions are far more the norm. Often they entail structural assumptions, such as a solvability condition, as well as necessary ones.

Representations of the structure $\mathcal{C} = \langle \Pi_i C_i, \succsim \rangle$ of commodity bundles are usually taken in economics to be n-tuples $\langle \varphi_1, \ldots, \varphi_n \rangle$ of functions, where φ_i maps C_i into $\mathbb{R}^+$, such that for each x_i and y_i in C_i, $i = 1, \ldots, n$,

$$(x_1,\ldots,x_n)\succsim(y_1,\ldots,y_n)\text{iff}\sum_i\varphi_i(x_i)\geq\sum_i\varphi_i(y_i). \tag{2}$$

In the measurement literature such a conjoint representation is called 'additive'.

Debreu, Luce and Tukey, Scott, Tversky, and others gave axioms about $\mathcal{C}$ for which existence of an additive representation can be shown, and such that any two representations $\langle\varphi_1, \ldots, \varphi_n\rangle$ and $\langle\psi_1, \ldots, \psi_n\rangle$ are related by affine transformations of the form $\psi_I = r\varphi_I + s_i$, $i = 1,\ldots, n$, $r > 0$. Note that r is common to all components. In Stevens' nomenclature, the set of such representations ψ_i for each fixed i are said to form an 'interval scale'.

In the example of the subjective probability structure, eq. (1), the usual sort of representation is a probability function P from $\mathcal{E}$ into [0, 1], such that, for all A, B in $\mathcal{E}$,

(i) $\mathrm{P}(\Omega) = 1$ and $\mathrm{P}(\varnothing) = 0$,
(ii) $A \succsim B$ iff $P(A) \geq P(B)$, and
(iii) if $A\cap B = \varnothing$, then $\mathrm{P}(A\cup B) = P(A) + P(B)$.

Unlike the previous two examples, here any two representations are identical, which scales Stevens called 'absolute'. Such a scale might be appropriate for representing a qualitative structure describing a relative frequency approach to probability. However, for subjective probability, it is better to view P as being a representation of the bounded ratio scale $\{rP|r > 0\}$ that is normalized by setting the bound, Ω, to be $1 = rP(\Omega)$.

A number of authors have given sufficient conditions in terms of the primitives for P to exist. Fine (1973) gave the first good, early summary of a variety of approaches to probability. Additional approaches to qualitative and subjective probability can be found in Narens (2008).

Interlocked Measurement Structures

A very common, and fundamentally important, feature of measurement is the existence of two or more ways to manipulate the same attribute. Again, mass measurement is illustrative. The mass order $\langle X, \succsim\rangle$ is determined as above. Mass can be manipulated in at least two ways by varying volumes and/or substances. Let $\langle V, \succsim', \mathrm{o}_V\rangle$ be a structure for combining volumes, where o_V is a set of volumes and V is a strictly monotonic, positive, and associative operation over V, and let $X = V \times S$ be a structure of masses, where S is a set of homogeneous substances of various densities. (v, s) is interpreted as an object of volume v filled with substance s and that, therefore, has mass. By definition, o_V is the operation on $V \times \{s\}$ such that (v,s) $\mathrm{o}_V(v',s) = (v\mathrm{o}_V v',s)$. The first manipulation is to vary $\succsim$ via volume concatenation of a single homogeneous material s, $\langle V \times \{s\}, \succsim, o_V\rangle$. The second is to manipulate the conjoint trade-off between volumes and substances, $\langle V \times S, \succsim\rangle$. Let m and m^* be the resulting representations of mass which, because they both preserve $\succsim$, must be strictly monotonically related. The ordering interlock alone is insufficient to develop measurement as was done in classical physics and as reflected in the familiar structure of physical units. Comparable developments are now beginning to appear in the behavioural and social sciences. The two structures must be interlocked beyond $\succsim$. Such interlocks are often types of distribution laws. In the mass case, the *distributive interlock* is: For $u, v \in V$ and $r, s \in S$,

$$(u,r)\sim(\mathrm{v},s)\text{ and }(u',r)\sim(v',s)\text{ imply }(u,\mathrm{o}_V u',r)\sim(u,\mathrm{o}_V v',s).$$

For much more detail, see Luce et al. (1990). Such laws are the source of the structure reflected in the units of physical measurement that are used and underlie dimensional analysis (Krantz et al. 1971; Luce et al. 1990; Narens 2002).

Typically, one is able to use the two separate numerical representations to reduce the interlock to solving a functional eq. (A functional equation resembles a differential one in that its solutions are the unknown functions satisfying the equation. It is unlike a differential equation in that no derivatives are involved; rather, the equation relates the value of the function at several values of the independent variable. See Aczél (1966, 1987) for a general introduction and classical examples of functional equations. Some arising in the behavioural and social sciences were novel and have required the aid of specialists to solve.)

Behavioural examples of interlocked structures are cited in sects. "Interlocked Structures and Applications to Utility Theory" and "Other Applications of Behavioural Interest".

Empirical Usefulness of Axiomatic Treatments

One, seemingly under-appreciated, advantage of a measurement approach to some scientific questions is that it offers an alternative way of testing quantitative models other than attempting to fit the representation to data and to evaluate it by a measure of goodness of fit. Because representations, such as utility and subjective probability, in general have free parameters and often free functions, estimation is necessary. In contrast, the axioms underlying such representations are (usually) parameter free. Testing the axioms often makes clear the source of a problem, thereby giving insight into what must be altered. Not everyone values the overall axiomatic (as compared with an analytic mathematical) approach to scientific questions; in particular, Anderson (1981, pp. 347–56) has sharply attacked it.

A familiar economic example arose in the theory of subjective expected utility (Fishburn 1970; Savage 1954). In its simplest form the domain is gambles of the form $x\mathrm{o}_A y$, meaning that x is the consequence attached to the occurrence of the chance event A, whereas y is the consequence when the chance outcome is $\neg A$. The x and y may be pure consequences or may be themselves gambles, and the theory postulates a preference ordering $\succsim$ over the pure consequences and gambles constructed from pure consequences and gambles. Classical axiomatizations establish conditions on preferences over gambles so that there exists a probability measure P on the algebra of events, as in a probability structure, and a 'utility function' U over the gambles such that U preserves $\succsim$ and

$$U(x\mathrm{o}_A y) = P(A)U(x) + [1 - P(A)]U(y). \quad (3)$$

A series of early empirical studies (for summaries see Allais and Hagen 1979; Kahneman and Tversky 1979) made clear that this representation, which can be readily defended on grounds of rationality, fails to describe human behaviour. Among its axioms, the one that appears to be the major source of difficulty is the 'extended sure-thing principle'. It may be stated as follows: For events A, B and C, with C disjoint from A and B,

$$x\mathrm{o}_A y \succsim x\mathrm{o}_B y \text{ iff } x\mathrm{o}_{A\cup C} y \succsim x\mathrm{o}_{B\cup C} y. \quad (4)$$

It is easy to verify that eq. (3) implies equation eq. (4), but people seem unwilling to abide by eq. (4). Any attempt at a descriptive theory must abandon it (see below).

Non-uniqueness of Axiom Systems

The isolation of properties in the axiomatic approach has an apparently happenstance quality because the choice of qualitative axioms is by no means uniquely determined by the representation. Any infinite structure has an infinity of equivalent axiom systems, and it is by no means clear why we select the ones that we do. It is entirely possible for a descriptive failure to be easily described in one axiomatization and to be totally obscure in another. Thus, some effort is spent on finding alternative but equivalent axiomatizations.

A related use of axiomatic methods, including the notion of scale (see sects. "Representations and Scales" and "Scale Types") is to study scientific meaningfulness, which is treated under meaningfulness and invariance.

Scale Types

Classification

As was noted in the examples, scale type has to do with the nature of the set of maps from one numerical representation of a structure into all other equally good representations, in a particular numerical structure such as the multiplicative real numbers. For some fixed numerical structure $\mathcal{R}$, a scale of the structure $\mathcal{X}$ is the collection of all $\mathcal{R}$-representations of $\mathcal{X}$. Much the simplest case, the one to which we confine most of our attention, occurs when $\mathcal{X}$ is totally ordered, the domain of $\mathcal{R}$ is either $\mathbb{R}$ or $\mathbb{R}^+$, and the $\mathcal{R}$-representations are all onto the domain and so are isomorphisms. Such scales are then usually described in terms of the

(mathematical) group of real transformations that take one representation into another. As Stevens (1946) noted, four distinct groups of transformations have appeared in physical measurement: any strictly monotonic function, any linear function $rx + s$, $r > 0$, any similarity transformation rx, $r > 0$, and the identity map. The corresponding scales are called *ordinal*, *interval*, *ratio*, and *absolute.* (Throughout this article, although not in all of the literature, ratio scales are assumed to be onto $\mathbb{R}^+$ thereby ruling out cases where an object maps to zero.)

A property of the first three scale types, called *homogeneity*, is that for each element x in the qualitative structure and each real number r in the domain of $\mathcal{R}$, some representation maps x into r. Homogeneity, which is typical of physical measurement, plays an important role in formulating many physical laws. Two general questions are: what are the possible groups associated with homogeneous scales, and what are the general classes of structures that can are represented by homogeneous scales?

It is easiest to formulate answers to these questions in terms of automorphisms (= symmetries), that is, isomorphisms of the qualitative structure onto itself. The representations and the automorphisms of the structure are in one-to-one correspondence, because, if φ and ψ are two representations and juxtaposition denotes function composition, then $\psi^{-1}\varphi$ is an automorphism of the structure, and if φ is a representation and α is an automorphism, then $\psi = \varphi\alpha$ is a representation.

It is not difficult to see that homogeneity of a scale simply corresponds to there being an automorphism that takes any element of the domain of the structure into any other element. To make this more specific, for M a positive integer, $\mathcal{X}$ is said to be *M-point homogeneous* if and only if each strictly ordered set of M points can be mapped by an automorphism onto any other strictly ordered set of M points. A structure that fails to be homogeneous for $M = 1$ is said to be *0-point homogeneous*; one that is homogeneous for every positive integer M is said to be *∞-point homogeneous*.

A second important feature of a scale is its degree of redundancy, formulated as follows: a scale is said to be *N-point unique*, where N is a non-negative integer if and only if for every two representations φ and ψ in the scale that agree at N distinct points, $\varphi = \psi$. By this definition, ratio scales are 1-point unique, interval scales are 2-point unique, and absolute scales 0-point unique. Scales, such as ordinal ones, that take infinitely many points to determine a representation are said to be *∞-point unique.* Equally, we speak of the structure being *N-point unique* if and only if every two automorphisms that agree at N distinct points are identical.

The abstract concept of scale type can be given in terms of these concepts. The *scale type* of $\mathcal{X}$ is the pair (M, N) such that M is the maximum degree of homogeneity and N is the minimum degree of uniqueness of $\mathcal{X}$. For the types of cases under consideration, it can be shown that $M \leq N$. Ratio scales are of type (1, 1) and interval scales of type (2, 2). Narens (1981a, b) showed that the converses of both statements are true. And Alper (1987) showed that, if $M > 0$ and $N < \infty$, then $N = 1$ or 2. The group in the (1, 2) case consists of transformations of the form $rx + s$, where s is any real number and r is in some non-trivial, proper subgroup of the multiplicative group $\langle \mathbb{R}^+, \cdot \rangle$, One example is $r = k^n$, where $k > 0$ is fixed and n ranges over the integers. So a structure is homogeneous if and only if it is of type (1, 1), (1, 2), (2, 2), or (M, ∞). The (M, ∞) case is not fully understood. Ordinal scalable (∞,∞) structures appear frequently in science, and a $(1, \infty)$ structure for threshold measurement appears in psychophysics. We focus here on the (1, 1), (1, 2), (2, 2) cases. For detailed references, see Luce et al. (1990), Narens (1985), or Narens (2007).

Unit Representations of Homogeneous Concatenation Structures

The next question is: which structures have scales of these types? Although the full answer is unknown, it is completely understood for ordered structures with binary operations. This is useful because, as was noted, the associative form of these operations plays a central role in much physical measurement and, as we shall see below, both associative and non-associative forms arise naturally in two distinct ways of interest to behavioural and social scientists.

Consider real concatenation structures of the form $\mathcal{R} = \langle \mathbb{R}^+, \geq, *' \rangle$ where $\geq$ has its usual meaning and we have replaced + by a general binary, numerical operation $*'$ that is strictly monotonic in each variable. The major result is that if $\mathcal{X}$ satisfies $M > 0$ and $N < \infty$ (a sufficient condition for finite N is that $*'$ be continuous – Luce and Narens 1985) then the structure can be mapped canonically into an isomorphic one of the form $\langle \mathbb{R}^+, \geq, * \rangle$, with a function f from $\mathbb{R}^+$ onto $\mathbb{R}^+$ such that

(i) f is strictly increasing,
(ii) $f(x)/x$ is strictly decreasing, and
(iii) for all x, y in $\mathbb{R}^+$, $x*y = yf(x/y)$ (Cohen and Narens 1979)

This type of canonical representation, which is called a *unit representation*, is invariant under the similarities of a ratio scale, that is, for each positive real r,

$$rx * ry = ryf(rx/ry) = ryf(x/y) = r(x * y).$$

The two most familiar examples of unit representations are ordinary additivity, for which $f(z) = 1 + z$ and so $x*y = x + y$, and bisymmetry, for which $f(z) = z^c, c \in (0, 1)$, and so $x*y = x^c y^{1-c}$. Situations where such representations arise are discussed later.

A simple invariance property of the function f corresponds to the three finite scale types (Luce and Narens 1985). Consider the values of $\rho > 0$ for which $f(x^\rho) = f(x)^\rho$ for all $x > 0$. The structure is of scale type (1, 1) if and only if $\rho = 1$; of type (1, 2) if and only if for some fixed $k > 0$ and all integers n, $\rho = k^n$; and of type (2, 2) if and only if there are constants c and d in (0,1) such that

$$f(z) = \begin{cases} z^c, & z \geq 1 \\ z^d, & z < 1. \end{cases}$$

If, as is the usual practice in the social sciences (see subjective expected utility, sect. "Interlocked Structures and Applications to Utility Theory"), but not in physics, the above representation is transformed by taking logarithms, it becomes a weighted additive form on $\mathbb{R}$:

$$x * y = \begin{cases} cx + (1 - c)y, & x \geq y \\ dx + (1 - d)y, & x < y. \end{cases}$$

That representation is called *dual bilinear* and the underlying structures are called dual bisymmetric (when $c = d$, the 'dual' is dropped). For references see Luce et al. (1990).

Axiomatization of Concatenation Structures

Given this understanding of the possible representations of homogeneous, finitely unique concatenation structures, it is natural to return to the classical question of axiomatizing the qualitative properties that lead to them. Until the 1970s, the only two cases that were understood axiomatically were those leading to additivity and averaging (see below). We now know more, but our knowledge remains incomplete.

Additive Representations

The key mathematical result underlying extensive measurement, due to O. Hölder, states that when a group operation and a total ordering interlock so that the operation is strictly monotonic and is Archimedean in the sense that sufficient copies of any positive element (that is, any element greater than the identity element) will exceed any fixed element, then the group is isomorphic to an ordered subgroup of the additive real numbers. Basically, the theory of extensive measurement restricts itself to the positive subsemigroup of such a structure. Extensive structures can be shown to be of scale type (1, 1).

Various generalizations involving partial operations (defined for only some pairs of objects) have been developed. (For a summary, see Krantz et al. 1971, chs. 2, 3, and 5; Luce et al. 1990, ch. 19). Not only are these structures with partial operations more realistic, they are essential to an understanding of the partial additivity that arises in such cases as probability structures. They can be shown to be of scale type (0, 1). Michell (1999) gives an alternative perspective on measurement in the behavioural sciences and a critique of axiomatic measurement approaches.

The representation theory for extensive structures not only asserts the existence of a numerical representation, but provides a systematic algorithm (involving the Archimedean property) for constructing one to any pre-assigned degree of accuracy. This construction, directly or indirectly, underlies the extensive scales used in practice.

The second classical case, due to J. Pfanzagl, leads to weighted average representations. The conditions are monotonicity of the operation, a form of solvability, an Archimedean condition, and bisymmetry, $(x \circ u) \circ (y \circ v) \sim (x \circ y) \circ (u \circ v)$ which replaces associativity. One method of developing these representations involves two steps: first, the bisymmetric operation is recoded as a conjoint one (see sect. "Axiomatization of Conjoint Structures") as follows: $(u, v) \succsim (x,y)$ iff $u \circ v \succsim x \circ y$; and second, the conjoint structure is recoded as an extensive operation on one of its components. This reduces the proof of the representation theorem to that of extensive measurement, that is to Hölder's theorem, and so it too is constructive.

Non-additive Representations

The most completely understood generalization of extensive structures, called positive concatenation structures or PCSs for short, simply drops the assumption of associativity. Narens and Luce (see Narens 1985; Luce et al. 1990, ch. 19) showed that this was sufficient to get a numerical representation and that, under a slight restriction which has since been removed, the structure is 1-point unique, but not necessarily 1-point homogeneous. Indeed, Cohen and Narens (1979) showed that the automorphism group is an Archimedean ordered group and so is isomorphic to a subgroup of the additive real numbers; it is homogeneous only when the isomorphism is to the full group. As in the extensive case, one can use the Archimedean axiom to construct representations, but the general case is a good deal more complex than the extensive one and almost certainly requires computer assistance to be practical.

For Dedekind complete PCSs that map onto $\mathbb{R}^+$, a nice criterion for 1-point homogeneity is that, for each positive integer n and every x and y, then $n(x \circ y) = nx \circ ny$, where by definition $1x = x$ and $nx = (n - 1)x \circ x$. The form of the representations of all such homogeneous representations was described earlier.

The remaining broad type of concatenation structures consists of those that are idempotent: that is, for all x, $x \circ x = x$. The following conditions have been shown to be sufficient for idempotent structures to have a numerical representation (Luce and Narens 1985): $\circ$ is an operation that is strictly monotonic and satisfies an Archimedean condition (for differences) and a solvability condition that says for each x and y, there exist u and v such that $u \circ x = y = x \circ v$. If, in addition, such a structure is Dedekind complete, it can be shown that it is N-point unique with $N \leq 2$.

Axiomatization of Conjoint Structures

Binary Structures

A second major class of measurement structures, widely familiar from both physics and the social sciences, comprises those involving two or more independent variables exhibiting a trade-off in the to-be-measured dependent variable. Their commonness and importance in physics is illustrated by familiar physical relations among three basic attributes, such as kinetic $= mv^2/2$, where m is the mass and v the velocity of a moving body. Such conjoint trade-off structures are equally common in the behavioural and social sciences: preference between commodity bundles or between gambles; loudness of pure tones as a function of signal intensity and frequency; trade-off between delay and amount of a reward, and so on. Although there is some theory for more than two independent variables in the additive case, with the general representation given by eq. (2), for present purposes we confine attention to the two-variable case $\langle X \times S, \succsim \rangle$. Michell (1990) gives detailed analyses of a number of behavioural examples.

As with concatenation structures, the simplest case to understand is the additive one in which the major non-structural properties are:

(i) *independence* (*monotonicity*): if $(x, s) \succsim (x', s)$ holds for some s, then it holds for all s in S, and the parallel statement for the other component.

Note that this property allows us to induce natural orderings, $\succsim_X$ on X and $\succsim_S$ on S;

(ii) *Thomsen condition*: if $(x, r) \sim (y, t)$ and $(y, s) \sim (z, r)$, then $(x, s) \sim (z, t)$; and

(iii) an *Archimedean condition* which says that if $\{x_i\}$ is a bounded sequence and if for some $r \succ s$ it satisfies $(x_i, r) \sim (x_{i+1}, s)$, then the sequence is finite. A similar statement holds for the other component.

These properties, together with some solvability in the structure, are sufficient to prove the existence of an interval scale, additive representation (for a summary of various results, see Krantz et al. 1971, chs 6, 7, and 9). The result has been generalized to non-additive representations by dropping the Thomsen condition, which leads to the existence of a non-additive numerical representation (Luce et al. 1990, chs 19 and 20). The basic strategy is to define an operation, say $\circ_X$ on component X, that captures the information embodied in the trade-off between components. The induced structure can be shown to consist of two PCSs pieced together at an element that acts like a natural zero of the concatenation structure. The results for PCSs are then used to construct the representation. As might be anticipated, $\circ_X$ is associative if and only if the conjoint structure satisfies the Thomsen condition.

Interlocked Structures and Applications to Utility Theory

Interlocked Conjoint/Extensive Structures

The next more complex structure has the form $\mathcal{D} = \langle X \times S, \succsim, \circ \rangle$, where $\circ$ is an operation on S. Such structures appear in the construction of the dimensional structure of physical units. The key qualitative axioms for physical measurement are that $\langle X \times S, \succsim \rangle$ is a conjoint structure satisfying independence, $\langle S, \succsim_S, \circ \rangle$ is an extensive structure, where $\succsim_s$ is the induced ordering on S, and $\mathcal{D}$ is distributive, that is,

$$\text{if } (x,p) \sim (y,q) \text{ and } (x,s) \sim (y,t), \text{then } (x, p\circ s) \sim (y, q\circ t).$$

These axioms yield the following representation for $\mathcal{D}$: There exists a ratio scale S for the extensive structure $\langle S, \succsim_S, \circ \rangle$ such that for each $\varphi \in S$ there exists ψ from X into the positive reals such that for all x, y in X and all s, t, p, q in S, there exists a representation φ on S that is part of a multiplicative representation of the conjoint structure and additive over the concatenation operation: that is,

(i) $(x, s) \succsim (y, t)$ iff $\psi(x)\varphi(s) \geq \psi(y)\varphi(t)$, and
(ii) $\varphi(p \circ q) = \varphi(p) + \varphi(q)$.

Discussions of how to construct the full algebra of physical dimensions using distributive structures and how to generalize these algebras to situations where there are no primitive associative operations are discussed in Luce et al. (1990) and Narens (2002).

Rationality Assumptions in Traditional Utility Theory

As was noted earlier, an extensive literature exists on preferences among uncertain alternatives, often called 'gambles'. The first major theoretical development was the axiomatization of subjective expected utility (SEU), which is a representation satisfying, in the binary case, eq. (3). Although such axiomatizations are defensible theories in terms of principles of rationality, they fail as descriptions of human behaviour. The rationality axioms invoked are of three quite distinct types.

First, preference is assumed to be transitive. This assumption has been shown to fail in various empirical contexts (especially multifactor ones), with perhaps the most pervasive and still ill-understood example being the 'preference reversal phenomenon', discovered by Slovic and Lichtenstein and investigated extensively by others, most famously by Grether and Plott 1979), and several later references given in Luce (2000, pp. 39–45). Nevertheless, transitivity is the axiom that is least easy to give up. Even subjects who violate it are not inclined to defend their 'errors'. A few attempts have been made to develop theories without it, but so far they are complex and have not received much empirical

scrutiny (Bell 1982; Fishburn 1982; 1985; Suppes et al. 1989, chs 16 and 17).

The second type of rationality postulates so-called 'accounting' principles in which two gambles are asserted to be equivalent in preference because when analysed into their component outcomes they are seen to be identical. For example, if xo_Ay is a gamble and $(xo_Ay)o_By$ means that if the event B occurs first and then, independent of it, A occurs, then on accounting grounds $(xo_Ay)o_By \sim (xo_By)o_Ay$ is rational because, on both sides, x is the outcome when A and B both occur (although in opposite orders) and y otherwise. One of the first 'paradoxes' of utility theory, that of Allais, is a violation of an accounting equation which assumes that certain probability calculations also take place.

The third type of rationality condition is the extended sure-thing principle, eq. (4). Its failure, which occurs regularly in experiments, is substantially the 'paradox' pointed out earlier by Ellsberg. Subjects have insisted on the reasonableness of their violations of this principle (MacCrimmon 1967).

Some Generalizations of SEU

Kahneman and Tversky (1979) proposed a binary modification of the expected utility representation designed to accommodate the last two types of violations, and Tversky and Kahneman (1992) generalized it to general finite gambles. During the 1980s and 1990s a great deal of attention was devoted to this general class of so-called rank- (and sometimes sign-) dependent representations (RDU or RSDU) (also called cumulative and Choquet 1953, representations). Summaries of this work, much of it of an axiomatic character for both risky cases, where probabilities are assumed known, and uncertain cases, where a subjective probability function is constructed, can be found in Quiggin (1993) and Luce (2000). These developments rests very heavily on modifying the distribution laws that are assumed. A far more general survey of utility theory, covering many aspects of it from an economic but not primarily an axiomatic measurement-theoretic perspective, is Barberà et al. (1998; 2004).

To return to an axiomatic approach, suppose in what follows that $x_1 \succsim x_2 \succsim \cdots \succsim x_n$ and their associated event partition is $(E_1, E_2, \ldots, E_n)$. Define $E(i) = \cup_{j=1}^{i} E_j$. The class of RDU representations involve proving from the axioms the existence of an order-preserving, utility function U over pure consequences and gambles and, in general, non-additive weighting function S over the chance events such that

$$U(x_1, E_1; x_2, E_2; \ldots; x_n, E_n) = \sum_{i=1}^{n} U(x_i)[S(E_i \cup E(i-1)) - S(E(i-1))]. \tag{5}$$

Note that the weighting function is essentially the incremental impact of adding E_i to $E(i-1)$. When S is finitely additive, that is, for disjoint A and B, $S(A \cup B) = S(A) + S(B)$, then eq. (5) reduces to subjective expected utility (SEU).

If there is a unique consequence e, sometimes called a reference level and sometimes taken to be no change from the status quo, then the consequences and gambles can be partitioned into gains, where $x_i \succsim e$, and the remainder, losses. In such cases, usually it follows from the assumptions made that $U(e) = 0$ and, usually, the weighting functions are sign dependent (that is, their form depends on whether their consequences are positive with respect to e or negative). Also, the RSDU representation includes cumulative prospect theory (Tversky and Kahneman 1992) as a special case having added restrictions on both U and S.

Other interesting developments involving different patterns of weighting are cited in Luce (2000).

A great deal of attention has been paid to issues of accounting for empirical phenomena discovered over the years that have discredited SEU and EU as descriptive models of human behaviour. For some summaries see Luce (2000) and Marley and Luce (2005). M.H. Birnbaum (numerous citations of his articles appear in the last reference) has discovered experimental designs that discredit a major feature of eq. (5) called *coalescing* or, equally, *event splitting:* Suppose $x_k = x_k = y$, then

$$(x_1, E_1; \ldots; y, E_k; y, E_{k+}1; \ldots; x_n, E_n) \sim (x_1, E_1; \ldots; y, E_k \cup E_{k+1}; \ldots; x_n, E_n). \quad (6)$$

The left-hand side of eq. (6) is called 'split' because y is attached to each of two events, E_k and E_{k+1}. The right-hand side is called 'coalesced' because y is attached to the single coalesced event $E_k \cup E_{k+1}$. Bimbaum has vividly demonstrated that experimental subjects often fail to split gambles in ways that help facilitate rational decisions. The other direction, coalescing, is effortless because no choice is involved. Indeed, Birnbaum (2007) has shown that splitting the branch (x_1, E_1), which has the best consequence, x_1, enhances the apparent worth of a gamble, whereas splitting (x_n, E_n), the branch with the poorest consequence, diminishes it. Long ago, he proposed a modified representation, called TAX, because it 'taxes' the poorest consequence in favour of the best one, which accommodates many empirical phenomena, including this one, but neither he nor anyone else has offered a measurement axiomatization of TAX. This remains an open problem.

Joint Receipt

Beginning in 1990, Luce and collaborators have investigated an operation $\bigoplus$ of joint receipt in gambling structures and ways that it may interlock with gambling structures. Its interpretation is suggested by its name, having two goods at once which, because $\bigoplus$ is assumed to be associative and commutative, can be extended to any finite number of goods. Several possible interlocking laws have been studied, and improved axiomatizations involving them have been given for a number of classical representations (for a summary, see Luce 2000). The representation that has arisen naturally is called p-additive (so named, at the suggestion of A.J. Marley, because it is the only polynomial form that can be transformed into an additive one), namely, for some real δ,

$$U(x \oplus y) = U(x) + U(y) + \delta U(x) U(y).$$

(By rescaling U there is no loss of generality in assuming that δ is either –1, 0, or 1.)

Lack of Idempotence and the Utility of Gambling

A feature of very many utility models, in particular, of all RDU or RSDU ones, is *idempotence*:

$$(x, E_1; \ldots; x, E_i; \ldots; x, En) \sim x.$$

Among other things, this has been thought to be a way to connect gambles to pure consequences, but that feature is redundant with the certainty principle $(x, E(n)) \sim x$. Further, if there is an inherent utility or disutility to risk or gambling, as widespread behaviour suggests there is – witness Las Vegas and mountain climbing – violations of idempotence assess it. Luce and Marley (2000) proposed partitioning a gamble g into a pure consequence, called a *kernel equivalent of g*, $KE(g)$, with the joint receipt, $\bigoplus$, of its unrewarded event structure (e, E_1; …; e, E_i; …; e, E_n), which is called an *element of chance*. Although they found properties of such a decomposition based on the assumption that utility is additive over joint receipt, $\bigoplus$, they did not discover much about the form of the utility of an element of chance. In the case of risk, further work has led to a detailed axiomatic formulation of that leads either to EU plus a Shannon entropy term, or to a linear weighted form plus entropy of some degree different from 1. In the case of uncertainty, the form for elements of risk is much less restrictive (Luce et al. 2008a, b). This risky form was first arrived at by Meginniss (1976) using a non-axiomatic approach. Because of the symmetry of entropy, this representation is unable to account for Birnbaum's differential event splitting data. This approach needs much more work.

Other Applications of Behavioural Interest

A Psychophysical One

A modified version of one of the RDU axiomatizations has been reinterpreted as a theory of global psychophysics, meaning that the focus is on the full dynamic range of intensity dimensions (for example, in audition the range 5–130 dB SPL; contemporary IRBs restrict the top of the range to

85 dB), not just local ranges as in discrimination studies. An example of the primitives are sound intensities x and y to the left and right ears, respectively, denoted (x, u), about which the respondent makes loudness judgments. Given two such stimuli, (x, x) and (y, y), $x > y$, and a positive number p, the respondent also can be requested to judge which stimulus (z, z) makes the subjective 'interval' from (y, y) to (z, z) seem to be p times the 'interval' from (y, y) to (x, x). The data are z, which we may denote in operator notation as $(x, x)\mathrm{o}_p(y,y) := (z, z)$. Luce (2002, 2004) (for a summary of theory, tests, and references, see Luce and Steingrimsson 2006) provided testable axioms, it is shown that there is a real valued mapping ψ, called a psychophysical function, and a numerical distortion function W such that

$$\Psi(x,u) = \Psi(x,0) + \Psi(0,u) + \delta\Psi(x,0)\Psi(0,u)(\delta \geq 0), \quad (7)$$

$$W(p) = \frac{\Psi\left[(x,x)\mathrm{o}_p(y,y)\right] - \Psi(y,y)}{\Psi(x,x) - \Psi(y,y)}(x > y \geq 0). \quad (8)$$

The axioms have been empirically tested by Steingrimsson and Luce in four papers. The 2005a focused on each structure, the conjoint one and the operator; the 2005b focused on the interlocks between them for audition. The results are supportive of the theory. Possible mathematical forms for Ψ and W have been reduced to testable conditions that, with one exception (the cases where $\delta \neq 0$, $\Psi(x,0)$ and $\Psi(0, x)$ are both power functions but with different exponents), were evaluated with considerable, but not perfect, support, for power functions (2006; 2007). Narens (1996) earlier proposed a closely related theory that included an axiom that forced $W(1) = 1$. Empirical data of Ellermeier and Faulhammer (2000) and Zimmer (2005) soundly rejected the joint hypothesis that W is a power function with $W(1) = 1$. Subsequent theory and experiments found considerable support for power functions with $W(1) \neq 1$.

Foundations of Probability

Today, the usual approach to probability theory is the classical one due to Kolmogorov (1933). It assumes that probability is a σ-additive (the countable extension of finite additivity) measure function P with a sure event having probability 1. It defines the important concepts of independence and conditional probability in terms of P.

There are many objections to this approach as a foundation for probability. A summary of most of them can be found in Fine (1973) and Narens (2008). In particular, independence and conditional probability appear to be more basic concepts than unconditional probability: for example, one often needs to know the independence of events in order to estimate probabilities. Also, in most empirical situations one cannot exactly pin down the probabilities: that is, there are many probability functions consistent with the data. This suggests that in such situations the underlying probabilistic concept should be a family of probability functions instead of a single probability function. Obviously, with many consistent probability functions explaining the data, the Kolmogorov method of defining independence by $P(A \cap B) = P(A)P(B)$ really does not work. These and other difficulties disappear with measurement-theoretic approaches to probability (for example, see Krantz et al. 1971; Fine 1973; Narens 1985, 2008). The qualitative approach provides richer and more flexible methods than Kolmogorov's for formulating and investigating the foundations of probability.

Both the Kolmogorov and the measurement-theoretic approaches assume an event space that is a Boolean algebra of subsets. This assumption works for most applications in science and is routinely assumed in theoretical and empirical studies of subjective probability. A major exception to it is quantum mechanics, where a different event space is needed (von Neumann 1995).

It is well-known that Boolean algebras of events correspond to the classical propositional calculus of logic. The classical propositional calculus captures deductions for propositions that are either true or false. It is not adequate for capturing various concepts of 'vagueness', 'ambiguity', or 'incompleteness based on lack of knowledge'. For these, logicians use nonclassical propositional calculi. In general, these nonclassical calculi cannot be interpreted as the classical propositional calculus

with 'true' and 'false' replaced with probabilities. It is plausible that some of the just-mentioned concepts are relevant to how individuals make judgements and decisions. Their incorporation into formal descriptions of behaviour requires the event space to be changed from the usual algebra of events used in the Kolmogorov approach to probability to a different kind of event space. This issue and proposals for alternative event spaces are discussed in detail in Narens (2008).

In summary, the Kolmogorov approach to probability is flawed at a foundational level and is too narrow to account for many important scientific phenomena. The measurement-theoretic approach is one alternative for providing a better foundation and generalizations for the kind of probability theory described by Kolmogorov. One should also consider the possibility of developing probabilistic theories for event spaces different from algebras of events, especially for phenomena that fall outside of usual forms of observation, including various phenomena arising from mentation.

See Also

- ▶ Expected Utility Hypothesis
- ▶ Meaningfulness and Invariance
- ▶ Measurement
- ▶ Non-expected Utility Theory
- ▶ Prospect Theory
- ▶ Savage's Subjective Expected Utility Model
- ▶ Utility

Bibliography

Aczél, J. 1966. *Lectures on functional equations and their applications*. New York/London: Academic Press.

Aczél, J. 1987. *A short course on functional equations: Based upon recent applications to the social and behavioral sciences*. Dordrecht: Reidel.

Alimov, N.G. 1950. On ordered semigroups. *Izvestia Akademii Nauk SSSR, Serija Mat.* 14: 569–576.

Allais, M., and O. Hagen. 1979. *Expected utility hypotheses and the allais paradox*. Dordrecht: Reidel.

Alper, T.M. 1987. A classification of all order-preserving homeomorphism groups that satisfy uniqueness. *Journal of Mathematical Psychology* 31: 135–154.

Anderson, N.H. 1981. *Foundations of information integration theory*. New York: Academic.

Barberà, S., P. Hammond and C. Seidl 1998, 2004. *Handbook of utility theory*, vol. I: *Principles*; vol. II: *Extensions*. Boston: Kluwer.

Bell, D. 1982. Regret in decision making under uncertainty. *Operations Research* 30: 961–981.

Birnbaum, M.H. 2007. Tests of branch splitting and branch-splitting independence in Allais paradoxes with positive and mixed consequences. *Organizational Behavior and Human Decision Processes* 102: 154–173.

Choquet, G. 1953. Theory of capacities. *Annals Institut Fourier* 5: 131–295.

Cohen, M., and L. Narens. 1979. Fundamental unit structures: A theory of ratio scalability. *Journal of Mathematical Psychology* 20: 193–232.

Ellermeier, W., and G. Faulhammer. 2000. Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception and Psychophysics* 62: 1505–1511.

Fine, T. 1973. *Theories of probability*. New York: Academic.

Fishbum, P.C. 1985. Nontransitive preference theory and the preference reversal phenomenon. *International Review of Economics and Business* 32: 39–50.

Fishburn, P.C. 1970. *Utility theory for decision making*. New York: Wiley.

Fishburn, P.C. 1982. Nontransitive measurable utility. *Journal of Mathematical Psychology* 26: 31–67.

Grether, D.M., and C.R. Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69: 623–638.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291.

Kolmogorov, A. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1946. New York: Chelsea.

Luce, R.D. 2000. *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Mahwah: Erlbaum. Errata: see Luce's web page at http://www.socsci.uci.edu.

Luce, R.D. 2002. A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review* 109: 520–532.

Luce, R.D. 2004. Symmetric and asymmetric matching of joint presentations. *Psychological Review* 111: 446–454.

Luce, R.D., and A.A.J. Marley. 2000. On elements of chance. *Theory and Decision* 49: 97–126.

Luce, R.D., and L. Narens. 1985. Classification of concatenation measurement structures according to scale type. *Journal of Mathematical Psychology* 29: 1–72.

Luce, R.D., and R. Steingrimsson. 2006. Global psychophysical judgments of intensity: Summary of a theory and experiments. In *Measurement and representations of sensations*, ed. H. Colonious and E. Dzharfov. Mahwah: Lawrence Erlbaum Associates.

Luce, R.D., D.H. Krantz, P. Suppes and A. Tversky 1990. *Foundations of measurement*, vol. 3. New York: Academic. Repr. New York: Dover, 2007.

Luce, R.D., C.T. Ng, J. Aczél, and A.A.J. Marley. 2008a. Utility of gambling I: Entropy-modified linear weighted utility. *Economic Theory* 36: 1–33.

Luce, R.D., C.T. Ng, J. Aczél, and A.A.J. Marley. 2008b. Utility of gambling II: Risk, paradoxes, data. *Economic Theory* 36: 165–187.

MacCrimmon, K.R. 1967. Descriptive and normative implications of the decision theory postulates. In *Risk and uncertainty*, ed. K. Borch and J. Mossin. New York: Macmillan.

Marley, A.A.J., and R.D. Luce. 2005. Independence properties vis-à-vis several utility representations. *Theory and Decision* 58: 77–143.

Meginniss, J.R. 1976. A new class of symmetric utility rules for gambles, subjective marginal probability functions, and a generalized Bayes' rule. *Proceedings of the American statistical association*, *business and economic statistics section*, 471–6.

Michell, J. 1990. *An introduction to the logic of psychological measurement*. Hillsdale: Lawrence Erlbaum Associates.

Michell, J. 1999. *Measurement in psychology: Critical history of a methodological concept*. Cambridge: Cambridge University Press.

Myung, J.I., G. Karabatsos, and G.J. Iverson. 2005. A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology* 49: 205–225.

Narens, L. 1981a. A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision* 13: 1–70.

Narens, L. 1981b. On the scales of measurement. *Journal of Mathematical Psychology* 24: 249–275.

Narens, L. 1985. *Abstract measurement theory.* Cambridge, MA: MIT Press.

Narens, L. 1996. A theory of ratio magnitude estimation. *Journal of Mathematical Psychology* 40: 109–129.

Narens, L. 2002. *Theories of meaningfulness*. Mahwah: Lawrence Erlbaum Associates.

Narens, L. 2007. *Introduction to the theories of measurement and meaningfulness and the use of invariance in science*. Mahwah: Lawrence Erlbaum Associates.

Narens, L. 2008. *Theories of probability: An examination of logical and qualitative foundations*. London: World Scientific.

Pfanzagl, J. 1968. *Theory of measurement.* New York: Wiley. 2nd edn, Vienna: Physica, 1971.

Quiggin, J. 1993. *Generalized expected utility theory: The rank-dependent model*. Boston: Kluwer.

Roberts, F.S. 1979. *Measurement theory.* Reading: Addison-Wesley.

Roberts, F.S., and R.D. Luce. 1968. Axiomatic thermodynamics and extensive measurement. *Synthese* 18: 311–326.

Savage, L.J. 1954. *The foundations of probability.* New York: Wiley.

Steingrimsson, R., and R.D. Luce. 2005a. Evaluating a model of global psychophysical judgments I: Behavioral properties of summations and productions. *Journal of Mathematical Psychology* 49: 290–306.

Steingrimsson, R., and R.D. Luce. 2005b. Evaluating a model of global psychophysical judgments II: Behavioral properties linking summations and productions. *Journal of Mathematical Psychology* 49: 308–319.

Steingrimsson, R., and R.D. Luce. 2006. Empirical Evaluation of a model of global psychophysical judgments III: A form for the psychophysical and perceptual filtering. *Journal of Mathematical Psychology* 50: 15–29.

Steingrimsson, R., and R.D. Luce. 2007. Empirical Evaluation of a model of global psychophysical judgments IV: Forms for the weighting function. *Journal of Mathematical Psychology* 51: 29–44.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science* 103: 677–680.

Suppes, P. 2002. *Representation and invariance of scientific structures*. Stanford: CSLI publications.

Suppes, P., Krantz, D.H., Luce, R.D. and Tversky, A. 1989. *Foundations of measurement*, vol. 2. New York: Academic. Repr. New York: Dover, 2007.

Tversky, A., and D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297–323.

von Neumann, J. 1995. *Mathematical foundations of quantum mechanics*. Princeton: Princeton University Press.

Zimmer, K. 2005. Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics* 67: 569–579.

Krantz, D.H., R.D. Luce, P. Suppes and A. Tversky 1971. *Foundations of measurement*, vol. 1. New York: Academic. Repr. New York: Dover, 2007.

Mechanism Design

Roger B. Myerson

Abstract

A mechanism is a specification of how economic decisions are determined as a function of the information that is known by the individuals in the economy. Mechanism theory shows that *incentive constraints* should be considered coequally with *resource constraints* in the formulation of the economic problem. Where individuals' private information and actions are difficult to monitor, the need to give people an incentive to share information and exert efforts may impose constraints on economic systems just as much as the limited

availability of raw materials. Mechanism design is the fundamental mathematical methodology for analysing economic efficiency subject to incentive constraints.

Keywords

Adverse selection; Bayesian collective-choice problems; Bayesian equilibrium; Bilateral bargaining; Coase theorem; Cooperative game theory; Correlated equilibrium; Direct-revelation mechanisms; First-price auctions; Incentive constraints; Incentive efficiency; Incentive-compatible mechanisms; Incomplete information; Inscrutability principle; Mechanism design; Pareto efficiency; Private information; Revelation principle; Revenue equivalence theorems; Sealed-bid auctions

JEL Classification

C7

Overview

A mechanism is a specification of how economic decisions are determined as a function of the information that is known by the individuals in the economy. In this sense, almost any kind of market institution or economic organization can be viewed, in principle, as a mechanism. Thus mechanism theory can offer a unifying conceptual structure in which a wide range of institutions can be compared, and optimal institutions can be identified.

The basic insight of mechanism theory is that *incentive constraints* should be considered coequally with *resource constraints* in the formulation of the economic problem. In situations where individuals' private information and actions are difficult to monitor, the need to give people an incentive to share information and exert efforts may impose constraints on economic systems just as much as the limited availability of raw materials. The theory of mechanism design is the fundamental mathematical methodology for analysing these constraints.

The study of mechanisms begins with a special class of mechanisms called *direct-revelation* mechanisms, which operate as follows. There is assumed to be a mediator who can communicate separately and confidentially with every individual in the economy. This mediator may be thought of as a trustworthy person, or as a computer tied into a telephone network. At each stage of the economic process, each individual is asked to report all of his private information (that is, everything that he knows that other individuals in the economy might not know) to the mediator. After receiving these reports confidentially from every individual, the mediator may then confidentially recommend some action or move to each individual. A direct-revelation mechanism is any rule for specifying how the mediator's recommendations are determined, as a function of the reports received.

A direct-revelation mechanism is said to be *incentive compatible* if, when each individual expects that the others will be honest and obedient to the mediator, then no individual could ever expect to do better (given the information available to him) by reporting dishonestly to the mediator or by disobeying the mediator's recommendations. That is, if honesty and obedience is an equilibrium (in the game- theoretic sense), then the mechanism is incentive compatible.

The analysis of such incentive-compatible direct-revelation mechanisms might at first seem to be of rather narrow interest, because such fully centralized mediation of economic systems is rare, and incentives for dishonesty and disobedience are commonly observed in real economic institutions. The importance of studying such mechanisms is derived from two key insights: (i) for any equilibrium of any general mechanism, there is an incentive-compatible direct-revelation mechanism that is essentially equivalent; and (ii) the set of incentive-compatible direct-revelation mechanisms has simple mathematical properties that often make it easy to characterize, because it can be defined by a set of linear inequalities. Thus, by analysing incentive-compatible direct-revelation mechanisms, we can characterize what can be accomplished in all possible equilibria of all possible mechanisms, for a given economic situation.

Insight (i) above is known as the *revelation principle.* It was first recognized by Gibbard (1973), but for a somewhat narrower solution concept (dominant strategies, instead of Bayesian equilibrium) and for the case where only informational honesty is problematic (no moral hazard). The formulation of the revelation principle for the broader solution concept of Bayesian equilibrium, but still in the case of purely informational problems, was recognized independently by many authors around 1978 (see Dasgupta et al. 1979; Harris and Townsend 1981; Holmstrom 1977; Myerson 1979; Rosenthal 1978). Aumann's (1974, 1987) concept of *correlated equilibrium* gave the first expression to the revelation principle in the case where only obedient choice of actions is problematic (pure moral hazard, no adverse selection). The synthesis of the revelation principle for general Bayesian games with incomplete information, where both honesty and obedience are problematic, was given by Myerson (1982). A generalization of the revelation principle to multistage games was stated by Myerson (1986).

The intuition behind the revelation principle is as follows. First, a central mediator who has collected all relevant information known by all individuals in the economy could issue recommendations to the individuals so as to simulate the outcome of any organizational or market system, centralized or decentralized. After the individuals have revealed all of their information to the mediator, he can simply tell them to do whatever they would have done in the other system. Second, the more information that an individual has, the harder it may be to prevent him from finding ways to gain by disobeying the mediator. So the incentive constraints will be least binding when the mediator reveals to each individual only the minimal information needed to identify his own recommended action, and nothing else about the reports or recommendations of other individuals. So, if we assume that the mediator is a discrete and trustworthy information-processing device, with no costs of processing information, then there is no loss of generality in assuming that each individual will confidentially reveal all of his information to the mediator (maximal revelation to the trustworthy mediator), and the mediator in return will reveal to each individual only his own recommended action (minimal revelation to the individuals whose behaviour is subject to incentive constraints).

The formal proof of the revelation principle is difficult only because it is cumbersome to develop the notation for defining, in full generality, the set of all general mechanisms, and for defining equilibrium behaviour by the individuals in any given mechanism. Once all of this notation is in place, the construction of the equivalent incentive-compatible direct-revelation mechanism is straightforward. Given any mechanism and any equilibrium of the mechanism, we simply specify that the mediator's recommended actions are those that would result in the given mechanism if everyone behaved as specified in the given equilibrium when his actual private information was as reported to the mediator. To check that this constructed direct-revelation mechanism is incentive compatible, notice that any player who could gain by disobeying the mediator could also gain by similarly disobeying his own strategy in the given equilibrium of the given mechanism, which is impossible (by definition of equilibrium).

Mathematical Formulations

Let us offer a precise general formulation of the proof of the revelation principle in the case where individuals have private information about which they could lie, but there is no question of disobedience of recommended actions or choices. For a general model, suppose that there are n individuals, numbered 1 to n. Let C denote the set of all possible combinations of actions or resource allocations that the individuals may choose in the economy. Each individual in the economy may have some private information about his preferences and endowments, and about his beliefs about other individuals' private information. Following Harsanyi (1967), we may refer to the state of an individual's private information as his *type.* Let T_i denote the set of possible types for any individual i, and let

$T = T_i \times \ldots \times T_n$ denote the set of all possible combinations of types for all individuals.

The preferences of each individual i may be generally described by some *payoff function* $u_i : C \times T \to \mathbb{R}$, where $u_i(c, (t_i, \ldots, t_n))$ denotes the payoff, measured in some von Neumann–Morgenstern utility scale, that individual i would get if c was the realized resource allocation in C when $(t_i, \ldots, t_n)$ denotes the actual types of the individuals 1, …, n 1,…,n respectively. For short, we may write $t = (t_i, \ldots, t_n)$ to describe a combination of types for all individuals.

The beliefs of each individual i, as a function of his type, may be generally described by some function $p_i(\cdot|\cdot)$, where $p_i, (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n|t_i)$ denotes the probability that individual i would assign to the event that the other individuals have types as in $(t_i, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$, when i knows that his own type is t_i. For short, we may write, $t_{-i}(t_i, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$, to describe a combination of types for all individuals other than i. We may let $T_{-i} = T_1 \times \ldots \times T_{i-1} \times \ldots \times T_n$ denote the set of all possible combinations of types for the individuals other than i.

The general model of an economy defined by these structures $(C, T_1, \ldots, T_n, u_1, \ldots, u_n, p_1, \ldots, p_n)$ is called a Bayesian collective-choice problem.

Given a Bayesian collective-choice problem, a general mechanism would be any function of the form $\gamma : S_1 \times \ldots \times S_n, \to C$, where, for each i, S_i is a nonempty set that denotes the set of strategies that are available for individual i in this mechanism. That is, a general mechanism specifies the strategic options that each individual may choose among, and the social choice or allocation of resources that would result from any combination of strategies that the individuals might choose. Given a mechanism, an equilibrium is any specification of how each individual may choose his strategy in the mechanism as a function of his type, so that no individual, given only his own information, could expect to do better by unilaterally deviating from the equilibrium. That is, $\sigma = (\sigma_1, \ldots, \sigma_n)$ is an equilibrium of the mechanism γ if, for each individual i, σ_i is a function from T_i to S_i, and, for every t_i in T_i and every s_i in S_i,

$$\Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(\gamma(\sigma(t)), t) \geq \Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(\gamma(\sigma_{-i}(t_{-i}), s_i), t).$$

(Here $\sigma(t) = (\sigma_1(t_1), \ldots, \sigma_n(t_n))$ and $(\sigma_{-i}(t_{-i}), s_i) = (\sigma_1(t_1), \ldots, \sigma_{i-1}(t_{i-1}), s_i, (\sigma_{i+1}(t_{i+1}), \ldots, \sigma_n(t_n))$.) Thus, in an equilibrium σ, no individual i, knowing only his own type t_i, could increase his expected payoff by changing his strategy from $\sigma_i(t_i)$ to some other strategy s_i, when he expects all other individuals to behave as specified by the equilibrium σ. (This concept of equilibrium is sometimes often called *Bayesian equilibrium* because it respects the assumption that each player knows only his own type when he chooses his strategy in S_i. For a comparison with other concepts of equilibrium, see Dasgupta et al. 1979, and Palfrey and Srivastava 1987.)

In this context, a direct-revelation mechanism is any mechanism such that the set S_i of possible strategies for each player i is the same as his set of possible types T_i. A direct-revelation mechanism is (Bayesian) incentive-compatible iff it is an equilibrium (in the Bayesian sense defined above) for every individual always to report his true type. Thus, $\mu : T_1 \times \ldots \times T_n \to C$ is an incentive-compatible direct- revelation mechanism if, for each individual i and every pair of types t_i and r_i in T_i,

$$\Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(\mu(t), t) \geq \Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(\mu(t_{-i}, r_i), t).$$

(Here $(t_{-i}, r_i) = (t_1, \ldots, t_i - 1, r_i, t_{i+1}, \ldots, t_n)$.) We may refer to these constraints as the *informational incentive constraints* on the direct-revelation mechanism μ. These informational incentive constraints are the formal representation of the economic problem of *adverse selection*, so they may also be called adverse-selection constraints (or self-selection constraints).

Now, to prove the revelation principle, given any general mechanism γ and any Bayesian equilibrium σ of the mechanism γ, let μ be the direct-revelation mechanism μ defined so that, for every t in T,

$$\mu(t) = \gamma(\sigma(t)).$$

M

Then this mechanism μ always leads to the same social choice as γ does, when the individuals behave as in the equilibrium σ. Furthermore, μ is incentive compatible because, for any individual i and any two types t_i and r_i in T_i,

$$\begin{aligned}
&\sum_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i)u_i(\mu(t), t) \\
&= \sum_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i)u_i(\gamma(\sigma(t)), t) \\
&\geq \sum_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i)u_i(\gamma(\sigma_{-i}(t_{-i}), \sigma_i(r_i)), t) \\
&= \sum_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i)u_i(\mu(t_{-i}, r_i), t).
\end{aligned}$$

Thus, μ is an incentive-compatible direct-revelation mechanism that is equivalent to the given mechanism γ with its equilibrium σ.

Notice that the revelation principle asserts that any pair consisting of a mechanism *and* an equilibrium is equivalent to an incentive-compatible direct- revelation mechanism. Thus, a general mechanism that has several equilibria may correspond to several different incentive-compatible mechanisms, depending on which equilibrium is considered.

Furthermore, the same general mechanism will generally have different equilibria in the context of different Bayesian collective-choice problems, where the structure of individuals' beliefs and payoffs are different. For example, consider a first-price sealed-bid auction where there are five potential bidders who are risk- neutral with independent private values drawn from the same distribution over \$0–\$10. If the bidders' values are drawn from a uniform distribution over this interval, then there is an equilibrium in which each bidder bids 4/5 of his value. On the other hand, if the bidders' values are drawn instead from a distribution with a probability density that is proportional to the square of the value, then there is an equilibrium in which each bidder bids 8/9 of his value. So in one situation the first-price sealed-bid auction (a general mechanism) corresponds to an incentive-compatible mechanism in which the bidder who reports the highest value gets the object for 4/5 of his reported value; but in the other situation it corresponds to an incentive-compatible mechanism in which the bidder who reports the highest value gets the object for 8/9 of his reported value. There is no incentive-compatible direct-revelation mechanism that is equivalent to the first-price sealed-bid auction in all situations, independently of the bidders' beliefs about each others' values. Thus, if we want to design a mechanism that has good properties in the context of many different Bayesian collective-choice problems, we cannot necessarily restrict our attention to incentive-compatible direct- revelation mechanisms, and so our task is correspondingly more difficult. (See Wilson 1985, for a remarkable effort at this kind of difficult question.)

Even an incentive-compatible mechanism itself may have other dishonest equilibria that correspond to different incentive-compatible mechanisms. Thus, when we talk about selecting an incentive-compatible mechanism and assume that it will then be played according to its honest equilibrium, we are implicitly making an assumption about the selection of an equilibrium as well as of a mechanism or communication structure. Thus, for example, when we say that a particular incentive- compatible mechanism maximizes a given individual's expected utility, we mean that, if you could choose any general mechanism for coordinating the individuals in the economy and if you could also (by some public statement, as a focal arbitrator, using Schelling's 1960, *focal-point effect)* designate the equilibrium that the individuals would play in your mechanism, then you could not give this given individual a higher expected utility than by choosing this incentive-compatible mechanism and its honest equilibrium.

In many situations, an individual may have a right to refuse to participate in an economic system or organization. For example, a consumer generally has the right to refuse to participate in any trading scheme and instead just consume his initial endowment. If we let $w_i(t_i)$ denote the utility payoff that individual i would get if he refused to participate when his type is t_i, and if we assume that an individual can make the choice not to participate after learning his type, then an incentive- compatible mechanism μ must also

satisfy the following constraint, for every individual i and every possible type t_i.

$$\Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(\mu(t), t) \geq w_i(t_i).$$

These constraints are called *participational incentive constraints,* or *individual- rationality* constraints.

In the analysis of Bayesian collective-choice problems, we have supposed that the only incentive problem was to get people to share their information, and to agree to participate in the mechanism in the first place. More generally, a social choice may be privately controlled by one or more individuals who cannot be trusted to follow some pre-specified plan when it is not in their best interests. For example, suppose now that the choice in C is privately controlled by some individual (call him 'individual 0') whose choice of an action in C cannot be regulated. To simplify matters here, let us suppose that this individual 0 has no private information. Let $p_0(t)$ denote the probability that this individual would assign to the event that $t = (t_1,...,t_n)$ is the profile of types for the other n individuals, and let $u_0(c, t)$ denote the utility payoff that this individual receives if he chooses action c when t is the actual profile of types. Then, to give this active individual an incentive to obey the recommendations of a mediator who is implementing the direct-revelation mechanism μ, μ must satisfy

$$\Sigma_{t \in T} p_0(t) u_0(\mu(t), t) \geq \Sigma_{t \in T} p_0(t) u_0(\delta(\mu(t)), t)$$

for every function $\delta : C \to C$. These constraints assert that obeying the actions recommended by the mediator is better for this individual than any disobedient strategy δ under which he would choose $\delta(c)$ if the mediator recommended c. Such constraints are called *strategic incentive constraints* or *moral-hazard constraints,* because they are the formal representation of the economic problem of moral hazard.

For a formulation of general incentive constraints that apply when individuals both have private information and control private actions, see Myerson (1982) or (1985).

Applications

In general, the mechanism-theoretic approach to economic problems is to list the constraints that an incentive-compatible mechanism must satisfy, and to try to characterize the incentive-compatible mechanisms that have properties of interest.

For example, one early contribution of mechanism theory was the derivation of general *revenue equivalence* theorems in auction theory. Ortega-Reichert (1968) found that, when bidders are risk-neutral and have private values for the object being sold that are independent and drawn from the same distribution, then a remarkably diverse collection of different auction mechanisms all generate the same expected revenue to the seller, when bidders use equilibrium strategies. In all of these different mechanisms and equilibria, it turned out that the bidder whose value for the object was highest would always end up getting the object, while a bidder whose value for the object was zero would never pay anything. By analysing the incentive constraints, Harris and Raviv (1981), Myerson (1981) and Riley and Samuelson (1981) showed that all incentive-compatible mechanisms with these properties would necessarily generate the same expected revenue, in such economic situations.

Using methods of constrained optimization, the problem of finding the incentive-compatible mechanism that maximizes some given objective (one individual's expected utility, or some social welfare function) can be solved for many examples. The resulting optimal mechanisms often have remarkable qualitative properties.

For example, suppose a seller, with a single indivisible object to sell, faces five potential buyers or bidders, whose private values for the object are independently drawn from a uniform distribution over the interval from \$0 to \$10. If the objective is to maximize the sellers' expected revenue, optimal auction mechanisms exist and all have the property that the object is sold to the bidder with the highest value for it, except that the seller keeps the object in the event that the bidders' values are all less than \$5. Such a result may seem surprising, because this event could occur

with positive probability (1/32) and in this event the seller is getting no revenue in an 'optimal' auction, even though any bidder would almost surely be willing to pay him a positive price for the object. Nevertheless, no incentive-compatible mechanism (satisfying the participational and informational incentive constraints) can offer the seller higher expected utility than these optimal auctions, and thus no equilibrium of any general auction mechanism can offer higher expected revenue either.

Maximizing expected revenue requires a positive probability of seemingly wasteful allocation.

The threat of keeping the object, when all bidders report values below $5, increases the seller's expected revenue because it gives the bidders an incentive to bid higher and pay more when their values are above $5. In many other economic environments, we can similarly prove the optimality of mechanisms in which seemingly wasteful threats are carried out with positive probability. People have intuitively understood that costly threats are often made to give some individual an incentive to reveal some information or choose some action, and the analysis of incentive constraints allows us to formalize this understanding rigorously.

In some situations, incentive constraints imply that such seemingly wasteful allocations may have to occur with positive probability in all incentive-compatible mechanisms, and so also in all equilibria of all general mechanisms. For example, Myerson and Satterthwaite (1983) considered bilateral bargaining problems between a seller of some object and a potential buyer, both of whom are risk-neutral and have independent private values for the object that are drawn out of distributions that have continuous positive probability densities over some pair of intervals that have an intersection of positive length. Under these technical (but apparently quite weak) assumptions, it is impossible to satisfy the participational and informational incentive constraints with any mechanism in which the buyer gets the object whenever it is worth more to him than to the seller. Thus, we cannot hope to guarantee the attainment of full *ex post* efficiency of resource allocations in bilateral bargaining problems where the buyer and seller are uncertain about each other's reservation prices. If we are concerned with welfare and efficiency questions, it may be more productive to try to characterize the incentive-compatible mechanisms that maximize the expected total gains from trade, or that maximize the probability that a mutually beneficial trade will occur. For example, in the bilateral bargaining problem where the seller's and buyer's private values for the object are independent random variables drawn from a uniform distribution over the interval from $0 to $10, both of these objectives are maximized subject to incentive constraints by mechanisms in which the buyer gets the object if and only if his value is greater than the seller's value by $2.50 or more. Under such a mechanism, the event that the seller will keep the object when it is actually worth more to the buyer has probability 7/32, but no equilibrium of any general mechanism can generate a lower probability of this event.

The theory of mechanism design has fundamental implications about the domain of applicability of Coase's (1960) theorem, which asserts the irrelevance of initial property rights to efficiency of final allocations. The unavoidable possibility of failure to realize mutually beneficial trades, in such bilateral trading problems with two-sided uncertainty, can be interpreted as one of the 'transaction costs' that limits the validity of Coase's theorem. Indeed, as Samuelson (1985) has emphasized, reassignment of property rights generally changes the payoffs that individuals can guarantee themselves without selling anything, which changes the right-hand sides of the participational incentive constraints, which in turn can change the maximal social welfare achievable by an optimal incentive-compatible mechanism.

For example, consider again the case where there is one object and two individuals who have private values for the object that are independent random variables drawn from a uniform distribution over the interval from $0 to $10. When we assumed above that one was the 'seller', we meant that he had the right to keep the object and pay nothing to anyone, until he agreed to some other arrangement. Now, let us suppose instead that the rights to the object are distributed equally between

the two individuals. Suppose that the object is a divisible good and each individual has a right to take half of the good and pay nothing, unless he agrees to some other arrangement. (Assume that, if an individual's value for the whole good is t_i, then his value for half would be $t_i/2$.) With this symmetric assignment of property rights, we can design incentive-compatible mechanisms in which the object always ends up being owned entirely by the individual who has the higher value for it, as Cramton et al. (1987) have shown.

For example, consider the game in which each individual independently puts money in an envelope, and then the individual who put more money in his envelope gets the object, while the other individual takes the money in both envelopes. This game has an equilibrium in which each individual puts into his envelope an amount equal to one-third of his value for the whole good. This equilibrium of this game is equivalent to an incentive-compatible direct-revelation mechanism in which the individual who reports the higher value pays one-third of his value to buy out the other individual's half-share. This mechanism would violate the participational incentive constraints if one individual had a right to the whole good (in which case, for example, if his value were \$10 then he would be paying \$3.33 under this mechanism for a good that he already owned). But with rights to only half of the good, no type of either individual could expect to do better (at the beginning of the game, when he knows his own value but not the other's) by keeping his half and refusing to participate in this mechanism.

More generally, redistribution of property rights tends to reduce the welfare losses caused by incentive constraints when it creates what Lewis and Sappington (1989) have called *countervailing incentives*. In games where one individual is the seller and the other is the buyer, if either individual has an incentive to lie, it is usually because the seller wants to overstate his value or the buyer wants to understate his value. In the case where either individual may buy the other's half-share, neither individual can be sure at first whether he will be the buyer or the seller (unless he has the highest or lowest possible value). Thus, a buyer-like incentive to understate values, in the event where the other's value is lower, may help to cancel out a seller-like incentive to overstate values, in the event where the other's value is higher.

The theory of mechanism design can also help us to appreciate the importance of mediation in economic relationships and transactions. There are situations in which, if the individuals were required to communicate with each other only through perfect noiseless communication channels (for example, in face-to-face dialogue), then the set of all possible equilibria would be much smaller than the set of incentive-compatible mechanisms that are achievable with a mediator. (Of course, the revelation principle asserts that the former set cannot be larger than the latter.)

For example, consider the following 'sender–receiver game' due to J. Farrell. Player 1 has a privately known type that may be α or β, but he has no payoff-relevant action to choose. Player 2 has no private information, but he must choose an action from the set $\{x, y, z\}$. The payoffs to players 1 and 2 respectively depend on 1's type and 2's action as follows.

	x	y	z
α	$2, 3$	$1, 2$	$0, 0$
β	$4, -3$	$8, -1$	$0, 0$

At the beginning of the game, player 2 believes that each of 1's two possible types has probability 1/2.

Suppose that, knowing his type, player 1 is allowed to choose a message in some arbitrarily rich language, and player 2 will hear player 1's message (with no noise or distortion) before choosing his action. In every equilibrium of this game, including the randomized equilibria, player 2 must choose y with probability 1, after every message that player 1 may choose in equilibrium (see Farrell 1993; Myerson 1988). If there were some message that player 1 could use to increase the probability of player 2 choosing x (for example, 'I am α, so choosing x would be best for us both!'), then he would always send such a message when his type was α. (It can be shown that no message could ever induce player 2 to randomize

between x and z.) So not receiving such a message would lead 2 to infer that 1's type was β, which implies that 2 would rationally choose z whenever such a message was not sent, so that both types of 1 should always send the message (any randomization between x and y is better than z for both types of 1). But a message that is always sent by player 1, no matter what his type is, would convey no information to player 2, so that 2 would rationally choose his *ex ante* optimal action y.

If we now allow the players to communicate through a mediator who uses a randomized mechanism, then we can apply the revelation principle to characterize the surprisingly large set of possible incentive-compatible mechanisms. Among all direct-revelation mechanisms that satisfy the relevant informational incentive constraints for player 1 and strategic incentive constraints for player 2, the best for player 2 is as follows: if player 1 reports to the mediator that his type is α then with probability 2/3 the mediator recommends x to player 2, and with probability 1/3 the mediator recommends y to player 2; if player 1 reports to the mediator that his type is β then with probability 2/3 the mediator recommends y to player 2, and with probability 1/3 the mediator recommends z to player 2. Notice that this mechanism is also better for player 1 than the unmediated equilibria when 1's type is α, although it is worse for 1 when his type is β.

Other mechanisms that player 2 might prefer would violate the strategic incentive constraint that player 2 should not expect to gain by choosing z instead of y when y is recommended. If player 2 could pre-commit himself always to obey the mediator's recommendations, then better mechanisms could be designed.

Efficiency

The concept of efficiency becomes more difficult to define in economic situations where individuals have different private information at the time when the basic decisions about production and allocation are made. A welfare economist or social planner who analyses the Pareto efficiency of an economic system must use the perspective of an outsider, so he cannot base his analysis on the individuals' private information. Otherwise, public testimony as to whether an economic mechanism or its outcome would be 'efficient' could implicitly reveal some individuals' private information to other individuals, which could in turn alter their rational behaviour and change the outcome of the mechanism! Thus, Holmstrom and Myerson (1983) argued that efficiency should be considered as a property of mechanisms, rather than of the outcome or allocation ultimately realized by the mechanism (which will depend on the individuals' private information).

Thus, a definition of Pareto efficiency in a Bayesian collective-choice problem must look something like this: 'a mechanism is efficient if there is no other feasible mechanism that may make some other individuals better off and will certainly not make other individuals worse off.' However, this definition is ambiguous in at least two ways.

First, we must specify whether the concept of feasibility takes incentive constraints into account or not. The concept of feasibility that ignores incentive constraints may be called *classical feasibility.* In these terms, the fundamental insight of mechanism theory is that incentive constraints are just as real as resource constraints, so that incentive compatibility may be a more fruitful concept than classical feasibility for welfare economics.

Second, we must specify what information is to be considered in determining whether an individual is 'better off' or 'worse off'. One possibility is to say that an individual is made worse off by a change that decreases his expected utility payoff as would be computed before his own type or any other individuals' types are specified. This is called the *ex ante* welfare criterion. A second possibility is to say that an individual is made worse off by a change that decreases his conditionally expected utility, given his own type (but not given the types of any other individuals). An outside observer, who does not know any individual's type, would then say that an individual may be made worse off, in this sense, if this conditionally expected utility were decreased for at least one possible type of the individual. This is called the *interim* welfare criterion. A third possibility is to

say that an individual is made worse off by a change that decreases his conditionally expected utility given the types of all individuals. An outside observer would then say that an individual may be worse off in this sense if his conditionally expected utility were decreased for at least one possible combination of types for all the individuals. This is called the *ex post* welfare criterion.

If each individual knows his own type at the time when economic plans and decisions are made, then the interim welfare criterion should be most relevant to a social planner. Thus, Holmstrom and Myerson (1983) argue that, for welfare analysis in a Bayesian collective-choice problem, the most appropriate concept of efficiency is that which combines the interim welfare criterion and the incentive-compatible definition of feasibility. This concept is called *incentive efficiency,* or *interim incentive efficiency.* That is, a mechanism $\gamma : T \to C$ is incentive efficient if it is an incentive-compatible mechanism and there does not exist any other incentive- compatible mechanism $\gamma : T \to C$ such that for every individual i and every type t_i in T_i,

$$\Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i)u_i(\gamma(t), t) \geq \Sigma_{t_{-i} \in T_{-i}} p_i(t_{-i})u_i((\mu(t), t),$$

and there is at least one type of at least one individual for which this inequality is strict. If a mechanism is incentive efficient, then it cannot be common knowledge among the individuals, at the stage when each knows only his own type, that there is some other incentive-compatible mechanism that no one would consider worse (given his own information) and some might consider strictly better.

For comparison, another important concept is classical *ex post* efficiency, defined using the *ex post* welfare criterion and the classical feasibility concept. That is, a mechanism $\mu : T \to C$ is *(classically) ex post efficient* iff there does not exist any other mechanism $\gamma : T \to C$ (not necessarily incentive compatible) such that, for every individual i and every combination of individuals' types t in $T = T_1 \times \ldots \times T_n$,

$$u_i(\gamma(t), t) \geq u_i((\mu(t), t),$$

with strict inequality for at least one individual and at least one combination of individuals' types.

The appeal of *ex post* efficiency is that there may seem to be something unstable about a mechanism that sometimes leads to outcomes such that, if everyone could share their information, they could identify another outcome that would make them all better off. However, we have seen that bargaining situations exist where no incentive-compatible mechanisms are *ex post* efficient. In such situations, the incentive constraints imply that rational individuals would be unable to share their information to achieve these gains, because if everyone were expected to do so then at least one type of one individual would have an incentive to lie.

Thus, a benevolent outside social planner who is persuaded by the usual Paretian arguments should choose some incentive-efficient mechanism. To determine more specifically an 'optimal' mechanism within this set, a social welfare function is needed that defines tradeoffs, not only between the expected payoffs of different individuals but also between the expected payoffs of different types of each individual. That is, given any positive utility-weights $\lambda_i(t_i)$ for each type t_i of each individual i, one can generate an incentive-efficient mechanism by maximizing

$$\sum_{i=1}^{n} \sum_{t_i \in T_i} \lambda_i(t_i) \sum_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i)u_i(\mu(t), t)$$

over all $\mu : T \to C$ that satisfy the incentive constraints; but different vectors of utility weights may generate different incentive-efficient mechanisms.

Bargaining Over Mechanisms

A positive economic theory must go beyond welfare economics and try to predict the economic institutions that may actually be chosen by the individuals in an economy. Having established that a social planner can restrict his attention to incentive- compatible direct-revelation mechanisms, which is a mathematically simple set, it is natural to assume that rational economic agents

who are themselves negotiating the structure of their economic institutions should be able to bargain over the set of incentive-compatible direct-revelation mechanisms. But if we assume that individuals know their types already at the time when fundamental economic plans and decisions are made, then we need a theory of mechanism selection by individuals who have private information.

When we consider bargaining games in which individuals can bargain over mechanisms, there should be no loss of generality in restricting our attention to equilibria in which there is one incentive-compatible mechanism that is selected with probability 1 independently of anyone's type. This proposition, called the *inscrutability principle,* can be justified by viewing the mechanism-selection process as itself part of a more broadly defined general mechanism and applying the revelation principle. For example, suppose that there is an equilibrium of the mechanism-selection game in which some mechanism μ would be chosen if individual 1's type were α and some other mechanism ν would be chosen if 1's type were β. Then there should exist an equivalent equilibrium of the mechanism-selection game in which the individuals always select a direct-revelation mechanism that coincides with mechanism μ when individual 1 confidentially reports type α to the mediator (in the implementation of the mechanism, after it has been selected), and that coincides with mechanism ν when 1 reports type β to the mediator.

However, the inscrutability principle does not imply that the possibility of revealing information during a mechanism-selection process is irrelevant. There may be some mechanisms that we should expect not to be selected by the individuals in such a process, precisely because some individuals would choose to reveal information about their types rather than let these mechanisms be selected. For example, consider the following Bayesian collective-choice problem, due to Holmstrom and Myerson (1983). There are two individuals, 1 and 2, each of whom has two possible types, α and β, which are independent and equally likely. There are three social choice options, called x, y and z. Each individual's utility for these options depends on his type according to the following table.

Option	1, α	1, β	2, α	2, β
x	2	0	2	2
y	1	4	1	1
z	0	9	0	−8

The incentive-efficient mechanism that maximizes the *ex ante* expected sum of the two individuals' utilities is as follows: if 1 reports type α and 2 reports α then choose x, if 1 reports type β and 2 reports α then choose z, and if 2 reports β then choose y (regardless of 1's report). However, Holmstrom and Myerson argue that such a mechanism would not be chosen in a mechanism-selection game that is played when 1 already knows his type, because, when 1 knows that his type is α, he could do better by proposing to select the mechanism that always chooses x, and 2 would always want to accept this proposal. That is, because 1 would have no incentive to conceal his type from 2 in a mechanism-selection game if his type were α (when his interests would then have no conflict with 2's), we should not expect the individuals in a mechanism-selection game to agree inscrutably to an incentive- efficient mechanism that implicitly puts as much weight on 1's type- β payoff as the mechanism described above.

For another example, consider again the sender–receiver game due to Farrell. Recall that y would be the only possible equilibrium outcome if the individuals could communicate only face-to-face, with no mediation or other noise in their communication channel. Suppose that the mechanism-selection process is as follows: first 2 proposes a mediator who is committed to implement some incentive- compatible mechanism; then 1 can either accept this mediator and communicate with 2 thereafter only through him, or 1 can reject this mediator and thereafter communicate with 2 only face-to-face. Suppose now that 2 proposes that they should use a mediator who will implement the incentive-compatible mediation plan that is best for 2 (recommending x with probability 2/3 and y with probability 1/3 if 1 reports α, recommending y with probability 2/3

and z with probability 1/3 if 1 reports β). We have seen that this mechanism is worse than y for 1 if his type is β. Furthermore, this mechanism would be worse than y for player 1 under the *ex ante* welfare criterion, when his expected payoffs for type α and type β are averaged, each with weight 1/2. However, it is an equilibrium of this mechanism-selection game for player 1 always to accept this proposal, no matter what his type is. If 1 rejected 2's proposed mediator, then 2 might reasonably infer that 1's type was β, in which case 2's rational choice would be z instead of y, and z is the worse possible outcome for both of 1's types.

Now consider a different mechanism-selection process for this example, in which the informed player 1 can select any incentive-compatible mechanism himself, with only the restriction that 2 must know what mechanism has been selected by 1. For any incentive-compatible mechanism μ, there is an equilibrium in which 1 chooses μ for sure, no matter what his type is, and they thereafter play the honest and obedient equilibrium of this mechanism. To support such an equilibrium, it suffices to suppose that, if any mechanism other than μ were selected, then 2 would infer that 1's type was β and therefore choose z. Thus, concepts like sequential equilibrium from non-cooperative game theory cannot determine the outcome of this mechanism-selection game, beyond what we already know from the revelation principle; we cannot even say that 1's selected mechanism will be incentive-efficient. To get incentive efficiency as a result of mechanism-selection games, we need some further assumptions, like those of cooperative game theory.

An attempt to extend traditional solution concepts from cooperative game theory to the problem of bargaining over mechanisms has been proposed by Myerson (1983, 1984a, b). In making such an extension, one must consider not only the traditional problem of how to define reasonable compromises between the conflicting interests of different individuals, but also the problem of how to define reasonable compromises between the conflicting interests of different types of the same individual. That is, to conceal his type in the mechanism-selection process, an individual should bargain for some inscrutable compromise between what he really wants and what he would have wanted if his type had been different; and we need some formal theory to predict what a reasonable inscrutable compromise might be. In the above sender–receiver game, where only type β of player 1 should feel any incentive to conceal his type, we might expect an inscrutable compromise to be resolved in favor of type α. That is, in the mechanism-selection game where 1 selects the mechanism, we might expect both types of 1 to select the incentive- compatible mechanism that is best for type α. (In this mechanism, the mediator recommends x with probability 0.8 and y with probability 0.2 if 1 reports α; and the mediator recommends x with probability 0.4, y with probability 0.4, and z with probability 0.2 if 1 reports β.) This mechanism is the *neutral optimum* for player 1, in the sense of Myerson (1983).

See Also

- ▶ Incentive Compatibility
- ▶ Mechanism Design Experiments
- ▶ Mechanism Design (New Developments)
- ▶ Revelation Principle

Bibliography

Aumann, R.J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67–96.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Cramton, P., R. Gibbons, and P. Klemperer. 1987. Dissolving a partnership efficiently. *Econometrica* 55: 615–632.

Dasgupta, P., P. Hammond, and E. Maskin. 1979. The implementation of social choice rules: Some general results on incentive compatibility. *Review of Economic Studies* 46: 185–216.

Farrell, J. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5: 514–531. Repr. in *Mathematical Models in Economics,* ed. M. Bacharach and M. Dempster. Oxford: Oxford University Press, 1997.

Gibbard, A. 1973. Manipulation of voting schemes: A general result. *Econometrica* 41: 587–602.

M

Harris, M., and A. Raviv. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49: 1477–1499.
Harris, M., and R.M. Townsend. 1981. Resource allocation under asymmetric information. *Econometrica* 49: 33–64.
Harsanyi, J.C. 1967. Games with incomplete information played by Bayesian players. *Management Science* 14: 159–182, 320–334, 481–502.
Holmstrom, B. 1977. *On incentives and control in organizations*. Ph.D. thesis, Graduate School of Business, Stanford University.
Holmstrom, B., and R.B. Myerson. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51: 1799–1819.
Lewis, T.R., and D.E.M. Sappington. 1989. Countervailing incentives in agency problems. *Journal of Economic Theory* 49: 294–313.
Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–74.
Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operation Research* 6: 58–73.
Myerson, R.B. 1982. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of Mathematical Economics* 10: 67–81.
Myerson, R.B. 1983. Mechanism design by an informed principal. *Econometrica* 51: 1767–1797.
Myerson, R.B. 1984a. Two-person bargaining problems with incomplete information. *Econometrica* 52: 461–487.
Myerson, R.B. 1984b. Cooperative games with incomplete information. *International Journal of Game Theory* 13: 69–86.
Myerson, R.B. 1985. Bayesian equilibrium and incentive compatibility. In *Social goals and social organization*, ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein. Cambridge: Cambridge University Press.
Myerson, R.B. 1986. Multistage games with communication. *Econometrica* 54: 323–358.
Myerson, R.B. 1988. Incentive constraints and optimal communication systems. In *Proceedings of the second conference on theoretical aspects of reasoning about knowledge*, ed. M.Y. Vardi. Los Altos: Morgan Kaufmann.
Myerson, R.B., and M. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.
Ortega-Reichert, A. 1968. *Models for competitive bidding under uncertainty*. Ph.D. thesis, Department of Operations Research, Stanford University.
Palfrey, T., and S. Srivastava. 1987. On Bayesian implementable allocations. *Review of Economic Studies* 54: 193–208.
Riley, J.G., and W.F. Samuelson. 1981. Optimal auctions. *American Economic Review* 71: 381–392.
Rosenthal, R.W. 1978. Arbitration of two-party disputes under uncertainty. *Review of Economic Studies* 45: 595–604.
Samuelson, W. 1985. A comment on the Coase Theorem. In *Game-theoretic models of bargaining*, ed. A.E. Roth. Cambridge: Cambridge University Press.
Schelling, T.C. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.
Wilson, R. 1985. Incentive efficiency of double auctions. *Econometrica* 53: 1101–1115.

Mechanism Design (New Developments)

Sandeep Baliga and Tomas Sjöström

Abstract

Mechanism design concerns the question: given some desirable outcome, can we design a game which produces it? This theory provides a foundation for many important fields, such as auction theory and contract theory. We survey the recent literature dealing with topics such as robustness of mechanisms, renegotiation and collusion. An important issue is whether simple and intuitively appealing mechanisms can be optimal. Finally, we discuss what can be learned from recent experiments.

Keywords

Adverse selection; Asymmetric information; Auctions; Centralization; Cheap talk; Collusion; Commitment; Consequentialism; Credibility; Decentralization; Delegation; Dominant strategy mechanisms; Free riding; Hold-up; Incentive compatibility; Incomplete contracts; Information rent; Limited liability; Mechanism design; Moral hazard; Nash equilibrium; Paretian liberal; Pooling equilibrium; Principal and agent; Prospect theory; Refinements of Nash equilibrium; Renegotiation; Repeated games; Revelation mechanism; Revelation principle; Side contracts; Side transfers; Signalling; Social choice functions; Social choice rules; Soft-budget constraint; Surplus; Truthful implementation

JEL Classifications
C7

Possibility Results and Robustness

Game theory provides methods to predict the outcome of a given game. Mechanism design concerns the *reverse* question: given some desirable outcome, can we design a game which produces it? Formally, the *environment* is $\langle A,N,\Theta\rangle$, where A is a set of feasible and verifiable alternatives or outcomes, $N = \{1,\ldots,n\}$ is a set of agents, and Θ is a set of possible *states of the world.* Except where indicated, we consider *private values* environments, where a state is $\theta = (\theta_1,\ldots,\theta_n) \in \times_i \Theta_i = \Theta$, each agent i knows his own 'type' $\theta_i \in \Theta_i$, and his payoff $u_t(a, \theta_i)$ depends only on the chosen alternative and his own type. (This does not rule out the possibility that the agents know something about each others' types.) If values are not private, then they are said to be *interdependent.* A *mechanism* or *contract* $\Gamma = (S,h)$ specifies a set of feasible actions S_i for each agent i, and an outcome function $h : S \equiv \times_{i=1}^{n} S_i \rightarrow A$. An outside party (a principal or social planner), or the agents themselves, want to design a mechanism which produces optimal outcomes. These are often represented by a *social choice rule* (SCR) $F: \Theta \rightarrow A$. A *social choice function* (SCF) is a single-valued SCR. Implicitly, it is assumed that the mechanism designer does not know the true θ, and this lack of information makes it impossible for her to directly choose an outcome in $F(\theta)$. Instead, she uses the more roundabout method of designing a mechanism which produces an outcome in $F(\theta)$, *whatever the true* θ *may be.*

In a *revelation mechanism*, each agent simply reports what he knows (so if agent i only knows θ_i then $Si = \Theta_i$). By definition, an *incentive compatible* revelation mechanism has a *truthful* Bayesian–Nash equilibrium, that is, it achieves *truthful implementation.* Truthful implementation plays an important role in the theory because of the revelation principle (see the dictionary entry on mechanism design, which surveys the early literature on truthful implementation). The early literature produced powerful results on optimal mechanisms for auction design, bargaining problems, and other applications. However, it generally made quite strong assumptions, for example, that the agents and the principal share a common prior over Θ, that the principal can *commit* to a mechanism, that the agents cannot side- contract and always use equilibrium strategies, and so on. We survey the recent literature which deals with these issues. In addition, we note that the notion of truthful implementation has a drawback: it does not rule out the possibility that non-truthful equilibria also exist, and these may produce suboptimal outcomes. (A non-truthful equilibrium may even Pareto dominate the truthful equilibrium for the agents, and hence provide a natural focal point for coordinating their actions.) To rule out the possibility of suboptimal equilibria, we may require *full implementation*: for all $\theta \in \Theta$, the set of equilibrium outcomes should precisely equal $F(\theta)$.

Maskin (1999) assumed *complete information:* each agent knows the true θ. If $n \geq 3$ agents know θ, then any SCF can be truthfully implemented: let the agents report θ, and if at least $n - 1$ agents announce the same θ then implement the outcome $F(\theta)$. Unilateral deviations from a consensus are disregarded, so truthtelling is a Nash equilibrium. Of course, this revelation mechanism will also have non-truthful equilibria. For full implementation, more complex mechanisms are required. (Even if $n = 2$, any SCF can be truthfully implemented if the principal can credibly threaten to 'punish' both agents if they report different states; in an economic environment, this might be achieved by making each agent pay a fine.)

A necessary condition for full Nash implementation is (*Maskin*) *monotonicity* (Maskin 1999). Intuitively, monotonicity requires that moving an alternative *up* in the agents' preference rankings should not make it *less* likely to be optimal. This condition can be surprisingly difficult to satisfy. For example, if the agents can have any complete and transitive preference relation on A, then any Maskin monotonic SCF must be a constant function (Saijo 1987). The situation is quite different if we consider *refinements* of Nash equilibrium.

For example, there is a sense in which almost any (ordinal) SCR can be fully implemented in *undominated Nash equilibrium* when the agents have complete information (Palfrey and Srivastava 1991; Jackson et al. 1994; Sjöström 1994). Chung and Ely (2003) showed that this possibility result is not robust to small perturbations of the information structure that violate private values (there is a small chance that agent i knows more about agent j's preferences than agent j does). The violation of private values is key. For example, in Sjöström's (1994) mechanism, an agent who knows his own preferences can eliminate his dominated strategies, and a second round of elimination of *strictly* dominated strategies generates the optimal outcome. This construction is robust to small perturbations that respect private values.

A different kind of robustness was studied by McLean and Postlewaite (2002). Consider an economic environment where each agent i observes an independently drawn signal t_i which is correlated with the state θ. The complete information structure is approximated by letting each agent's signal be very accurate. With complete information, any SCF can be truthfully implemented. McLean and Postlewaite (2002) show robustness to perturbations of the information structure: any outcome can be approximated by an incentive-compatible allocation, if the agents' signals are accurate enough. There is no need to assume private values.

The literature on *Bayesian mechanism design* typically assumes each agent i knows only his own type $\theta_i \in \Theta_i$, the agents share a common prior p over $\Theta \equiv \times_{i=1}^{n} \Theta_i$, and the principal knows p. In fact, for truthful implementation with $n \geq 3$, the assumption that the principal knows p is redundant. Suppose for any common prior p on Θ, there is an incentive-compatible revelation mechanism $\Gamma_p = \left(\times_{i=1}^{n} \Theta_i, h_p\right)$ By definition, Γ_p truthfully implements the SCF $F_p \equiv h_p$. The mechanism Γ_p is 'parametric', that is, it depends on p. To be specific, consider a quasi-linear public goods environment with independent types, and suppose Γ_p is the well-known mechanism of d'Aspremont and Gérard-Varet (1979). Now consider a nonparametric mechanism Γ, where each agent i announces p and θ_i. If at least $n - 1$ agents report the same p, the outcome is $h_p(\theta_1, \ldots, \theta_n)$. Now, if agent i thinks everyone will announce p truthfully, he may as well do so. If in addition he thinks the other agents report θ_{-i} truthfully, then he should announce θ_i truthfully by incentive compatibility of Γ_p. Therefore, for any common prior p, the nonparametric mechanism Γ truthfully implements F_p. In this sense, the principal can use Γ to extract the agents' shared information about p. Of course, this particular mechanism also has non-truthful equilibria. Choi and Kim (1999) *fully* implemented the d'Aspremont and Gérard-Varet (1979) outcome in *undominated* Bayesian–Nash equilibrium, using a nonparametric mechanism. Naturally, their mechanism is quite complex. Suppose we restrict attention to mechanisms where each agent i only reports θ_i, truthfully in equilibrium. Then the necessary and sufficient condition for full nonparametric Bayesian–Nash implementation for any common prior p is (dominant strategy) incentive compatibility plus the *rectangular property* (Cason et al. 2006).

The d'Aspremont and Gérard-Varet (1979) mechanism is budget balanced and surplus maximizing. The above argument shows that such outcomes can be truthfully implemented by a nonparametric mechanism in quasi-linear environments with independent types. As is well known, this cannot be achieved by any dominant strategy mechanism. Thus, in general, nonparametric truthful implementation is easier than dominant strategy implementation. However, there are circumstances where the two concepts coincide. Bergemann and Morris (2005a) consider a model where each agent i has a *payoff type* $\theta_i \in \Theta_i$ and a *belief type* π_i The payoff type determines the payoff function $u_i(a, \theta_i)$ while the belief type determines beliefs over other agents' types. The set of socially optimal outcomes $F(\theta)$ depends on payoff types, but not on beliefs. Bergemann and Morris (2005a) show that in quasi-linear environments with no restrictions on side payments (hence no budget-balance requirement), truthful implementation for all possible type spaces with a common prior implies dominant strategy implementation. (For related results, see section "Other Theoretical Issues".)

Bergemann and Morris (2005b) consider *full* implementation of SCFs in a similar framework. The SCF $F: \Theta \rightarrow A$ is *fully robustly implemented* if there exists a mechanism which fully implements F on all possible type spaces. They make no common prior assumption. Full robust implementation turns out to be equivalent to implementation using iterated elimination of strictly dominated strategies. Although a demanding concept, there are situations where full robust implementation is possible. For example, a Vickrey-Clarke-Groves (VCG) mechanism in a public goods economy with private values and strictly concave valuation functions achieves implementation in strictly dominant strategies. However, Bergemann and Morris (2005b) show the impossibility of full robust implementation when values are sufficiently interdependent.

A generalization of Maskin monotonicity called *Bayesian monotonicity* is necessary for ('parametric') full Bayesian–Nash implementation (Postlewaite and Schmeidler 1986; Palfrey and Srivastava 1989a; Jackson 1991). Again, refinements lead to possibility results (Palfrey and Srivastava 1989b). Another way to expand the set of implementable SCRs is *virtual* implementation (Abreu and Sen 1991; Duggan 1997). Serrano and Vohra (2001) argue that the sufficient conditions for virtual implementation are in fact quite strong.

The work discussed so far is *consequentialist:* only the final outcome matters. The mechanisms are clearly not meant to be descriptive of real-world institutions. For example, they typically require the agents to report 'all they know' before any decision is reached, an extreme form of centralized decision making hardly ever encountered in the real world. (The question of how much information must be transmitted in order to implement a given SCR is addressed by Hurwicz and Reiter 2006, and Segal 2004.) Delegating the power to make (verifiable) decisions to the agents would only create additional 'moral hazard' constraints, as discussed in the entry on mechanism design. Since centralization eliminates these moral hazard constraints, it typically strictly dominates decentralization in the basic model. However, as discussed below, by introducing additional aspects such as renegotiation and collusion, we can frequently prove the optimality of more realistic decentralized mechanisms. The implicit assumption is that decentralized decision making is in itself a good thing, which is a mild form of non-consequentialism. (Other non-consequentialist arguments are discussed in section "Other Theoretical Issues".) We might add that there is, of course, no way to eliminate the moral hazard constraints if the agents take *unverifiable* decisions that cannot be contracted upon. In this case, the issue of centralization versus decentralization of decisions is moot.

Renegotiation and Credibility

Suppose $n = 2$ and both agents know the true θ. If a revelation mechanism is used and the agents announce different states, then we cannot identify a deviator from a 'consensus', so it may be necessary to punish *both* agents in order to support a truth- telling equilibrium. But this threat is not credible if the agents can avoid punishment by renegotiating the outcome. Maskin and Moore (1999) capture the renegotiation process by an exogenously given function $r : A \times \Theta \rightarrow A$ which maps outcome a in state θ into an efficient outcome $r(a, \theta)$. They derive an incentive-compatibility condition which is necessary for truth-telling when $n = 2$, and show that *renegotiation monotonicity* is necessary for full Nash implementation (see also Segal and Whinston 2002).

The idea that renegotiation may preclude the implementation of the first-best outcome, even when information is complete, has received attention in models of bilateral trade with relationship-specific investments (the hold-up problem). It is possible to implement the first-best outcome if trade is one-dimensional and investments are 'selfish', in the sense that each agent's investment does not directly influence the other agent's payoff (Nöldeke and Schmidt 1995; Edlin and Reichelstein 1996). If investments are not selfish, then the first-best cannot always be achieved, while the second-best can often be implemented without any explicit contract

(Che and Hausch 1999). Segal (1999) found a similar result in a model with k goods and selfish investments, for k large (see also Maskin and Tirole 1999; Hart and Moore 1999). It should be noted that the case $n = 2$ is quite special, and adding a third party often alleviates the problem of renegotiation (Baliga and Sjöström 2006).

Credibility and renegotiation also impact trading with asymmetric information. Suppose the seller can produce goods of different quality, but the buyer's valuation is his private information. It is typically second-best optimal for the seller to offer a contract such that low-valuation buyers consume less than first-best quality ('underproduction'), while high-valuation buyers enjoy 'information rents'. Incentive compatibility guarantees that the buyer reveals his true valuation. Now suppose trading takes place twice, and the buyer's valuation does not change. Suppose the seller cannot credibly commit to a long-run (two-period) contract. If the buyer reveals his true valuation in the first period, then in the second period the seller will leave him no rent. This is typically not the second-best outcome. The seller may prefer a 'pooling' contract which does not fully reveal valuations in the first period, a commitment device which limits his ability to extract second period rents. This idea has important applications. When a regulator cannot commit to a long-run contract, a regulated firm may hide information or exert less effort to cut costs, the *ratchet effect* (Freixas et al. 1985). A borrower may not exert effort to improve a project knowing that a lender with deep pockets will bail him out, the *soft budget constraint* (Dewatripont and Maskin 1995a). These problems are exacerbated if the principal is well informed and cannot commit not to use his information. Institutional or organizational design can alleviate the problems. By committing to acquire less information via 'incomplete contracts', or by maintaining an 'arm's-length relationship', the principal can improve efficiency (Dewatripont and Maskin 1995b; Crémer 1995). Less frequent regulatory reviews offset the ratchet effect, and a decentralized credit market helps to cut off borrowers from future funding. Long-run contracts can help, but they may be vulnerable to renegotiation (Dewatripont 1989). In particular, the second-period outcome may be renegotiated if quality levels are known to be different from the first-best. Again, some degree of pooling may be optimal.

If the principal cannot commit even to short-run contracts, then, after receiving the agents' messages, she always chooses an outcome that is optimal given her beliefs. She cannot credibly threaten punishments that she would not want to carry out. Refinements proposed in the cheap-talk literature suggest that a putative pooling equilibrium may be destroyed if an agent can reveal information by 'objecting' in a credible way. This leads to a necessary condition for full implementation with complete information which is reminiscent of Maskin monotonicity, but which involves the principal's preferences (Baliga et al. 1997).

Collusion

A large literature on collusion was inspired by Tirole (1986). A key contribution was made by Laffont and Martimort (1997), who assumed an uninformed third party proposes side contracts. This circumvents the signalling problems that might arise if a privately informed agent makes collusive proposals. A side contract for a group of colluding agents is a *collusive mechanism* which must respect incentive compatibility, individual rationality and feasibility constraints. The original mechanism Γ, designed by the principal, is called the *grand mechanism.* The objective is to design an optimal grand mechanism when collusion is possible. Typically, collusion imposes severe limits on what can be achieved.

Baliga and Sjöström (1998) study a model with moral hazard and limited liability. Two agents share information not known to the principal: agent 1's effort is observed by both agents. Agent 2's effort is known only to himself. In the absence of collusion, the optimal grand mechanism specifies a 'message game': agent 2 reports agent 1's effort to the principal. Now suppose the agents can side contract on agent 1's effort, but not on agent 2's effort (which is unobserved). Side contracts can

specify side transfers as a function of realized output, but must respect limited liability. This collusion may destroy centralized 'message games', and we obtain a theory of optimal delegation of decision making. For some parameters, it is optimal for the principal to contract only with agent 2, and let agent 2 subcontract with agent 1. This is intuitive, since agent 2 observes agent 1's effort and can contract directly on it. More surprisingly, there are parameter values where it is better for the principal to contract only with agent 1.

Mookherjee and Tsumagari (2004) study a similar model, but with adverse selection: the agents privately observe their own production costs. In this model, delegating to a 'prime supplier' creates 'double marginalization of rents': the prime supplier uses underproduction to minimize the other agent's information rent. A centralized contract avoids this problem. Hence, in this model delegation is always strictly dominated by centralization, even though the agents can collude.

Mookherjee and Tsumagari (2004) assume the agents can side contract before deciding to participate in the grand contract. Che and Kim (2006) assume side contracting occurs only after the decision to participate in the grand mechanism has been made. In this case, collusion does not limit what the principal can achieve. Hence, the timing of side contracting is important. In a complete information environment with $n \geq 3$, Sjöström (1999) showed that neither renegotiation nor collusion limit the possibility of undominated Nash implementation.

Other Theoretical Issues

In quasi-linear environments with uncorrelated types, there exist incentive- compatible mechanisms which maximize the social surplus (for example, d'Aspremont and Gérard-Varet 1979). But the principal cannot extract all the surplus: the agents must get informational rents. However, Crémer and McLean (1988) showed that the principal can extract all the surplus in auctions with *correlated types*. McAfee and Reny (1992) extended this result to general quasi-linear environments.

Jehiel and Moldovanu (2001) considered a quasi-linear environment with multidimensional (uncorrelated) types and interdependent values. Generically, a standard revelation mechanism cannot be designed to extract information about multidimensional types, and no incentive-compatible and surplus-maximizing mechanism exists. Mezzetti (2004) presents an ingenious *two-stage* mechanism which maximizes the surplus in interdependent values environments, even when types are independent and multidimensional. In the first stage, the mechanism specifies an outcome decision but not transfers. Transfers are determined in the second stage by reports on payoffs realized by the outcome decision. Mezzetti (2007) shows that the principal can sometimes extract all the surplus by this method, even if types are independent. For optimal mechanisms for a profit-maximizing monopolist when consumers have multidimensional types and private values, see Armstrong (1996).

Incentive compatibility does not require that each agent has a dominant strategy. Nevertheless, incentive-compatible outcomes can often be replicated by dominant strategy mechanisms (Mookherjee and Reichelstein 1992). In quasi-linear environments, all incentive-compatible mechanisms that maximize the social surplus are *payoff-equivalent* to dominant strategy (VCG) mechanisms (Krishna and Perry 1997; Williams 1999). However, as pointed out above, dominant strategies (but not incentive compatibility) rules out budget balance.

Bergemann and Välimäki (2002) assume agents can update a common prior by costly information acquisition. Suppose a single-unit auction has two bidders i and j who observe statistically independent private signals θ_i and θ_j. Bidder i's valuation of the good is $u_i(\theta_i, \theta_j) = \alpha\theta_i + \beta\theta_j$, where $\alpha > \beta > 0$. Thus, values are interdependent. Efficiency requires that bidder i gets the good if and only if $\theta_i \geq \theta_j$. Suppose bidders report their signals, the good is allocated efficiently given their reports, and the winning bidder i pays the price $(\alpha + \beta)\,\theta j$. This VCG mechanism is incentive compatible (Maskin 1992). If bidder i acquires negative information which causes him to lose the auction, then he imposes a negative

externality on the other bidder (as $\beta > 0$). This implies the bidders have an incentive to collect too much information. Conversely, there is an incentive to collect too little information when $\beta < 0$. Bergemann and Välimäki (2002) provide a general analysis of these externalities. Similar externalities occur when members of a committee must collect information before voting. If the committee is large, each vote is unlikely to be pivotal, and free riding occurs. Persico (2004) shows how the optimal committee is designed to encourage the members to collect information.

Some authors reject consequentialism and instead emphasize agents' *rights.* For example, suppose a mechanism implements *envy-free outcomes*. An agent might still feel unfairly treated if his own bundle is worse than a bundle which another agent *had the right to choose* (but did not). Such agents may demand 'equal rights' (Gaspart 1995). Unfortunately, once we leave the classical exchange economy, Sen's 'Paretian liberal' paradox (Sen 1970) suggests that rights are incompatible with efficiency (Deb et al. 1997). Sen originally considered rights embodied in SCRs rather than mechanisms. Peleg and Winter (2002) study *constitutional implementation* where the mechanism embodies the same rights as the SCR it implements.

Learning from Experiments

Cabrales et al. (2003) tested the so-called canonical mechanism for Nash implementation. A Nash equilibrium was played only 13 per cent of the time (20 per cent when monetary fines were used). Remarkably, the optimal outcome was implemented 68 per cent of the time (80 per cent with 'fines'), because deviations from equilibrium strategies frequently did not affect the outcome. This suggests that a desirable property of a mechanism is *fault-solerance:* it should produce optimal outcomes even if some 'faulty' players deviate from the theoretical predictions. Eliaz (2002) showed that, if at most $k < \frac{1}{2} n - 1$ players are 'faulty' (that is, unpredictable), then full Nash implementation is possible if *no-veto-power* and $(k + 1)$*-monotonicity hold.*

Equilibrium play can be justified by epistemic or dynamic theories. According to epistemic theories, common knowledge about various aspects of the game implies equilibrium play even in one-shot games. Experiments provide little support for this. However, there is evidence that players can reach equilibrium through a dynamic adjustment process. If a game is played repeatedly, with no player knowing any other player's payoff function, the outcome frequently converges to a Nash equilibrium of the one-shot complete information game (Smith 1979). Dynamic theories have been applied to the mechanism design problem (for example, Cabrales and Ponti 2000). Chen and Tang (1998) and Chen and Gazzale (2004) argue that mechanisms which induce supermodular games produce good long-run outcomes. Unfortunately, these convergence results are irrelevant for decisions that are taken infrequently, or if the principal is too impatient to care only about the long-run outcome.

The idea of dominant strategies is less controversial than Nash equilibrium, and should be more relevant for decisions that are taken infrequently. Unfortunately, experiments on dominant-strategy mechanisms have yielded negative results. Attiyeh, Franciosi and Isaac (2000, p. 112) conclude pessimistically, 'we do not believe that the pivot mechanism warrants further practical consideration. . . . This is due to the fundamental failure of the mechanism, in our laboratory experiments, to induce truthful value revelation.' However, VCG mechanisms (such as the pivotal mechanism) frequently have a multiplicity of Nash equilibria, some of which produce suboptimal outcomes. Cason et al. (2006) did experiments with *secure* mechanisms, which fully implement an SCR both in dominant strategies and in Nash equilibria. The players were much more likely to use their dominant strategies in secure than in non-secure mechanisms. In the non-secure mechanisms, deviations from dominant strategies tended to correspond to Nash equilibria. However, these deviations typically did not lead to suboptimal *outcomes*. In this sense, the non- secure mechanisms were

fault-tolerant. Kawagoe and Mori (2001) report experiments where deviations from dominant strategies typically corresponded to suboptimal Nash equilibria.

In experiments, subjects often violate standard axioms of rational decision making. Alternative theories, such as prospect theory, fit the experimental evidence better. But, if we modify the axioms of individual behaviour, the optimal mechanisms will change. Esteban and Miyagawa (2005) assume the agents have Gul–Pesendorfer preferences (Gul and Pesendorfer 2001). They suffer from 'temptation', and may prefer a smaller menu (choice set) to a larger one. Suppose each agent first chooses a menu, and then chooses an alternative from this menu. Optimal menus may contain 'tempting' alternatives which are never chosen in equilibrium, because this relaxes the incentive-compatibility constraints pertaining to the choice of menu. Eliaz and Spiegler (2006) assume some agents are 'sophisticated' and some are 'naive'. Sophisticated agents know that they are dynamically inconsistent, and would like to commit to a future decision. Naive agents are unaware that they are dynamically inconsistent. The optimal mechanism screens the agents by providing commitment devices that are chosen only by sophisticated agents.

Experiments reveal the importance of human emotions such as spite or kindness (Andreoni 1995; Saijo 2003). In many mechanisms in the theoretical literature, by changing his strategy an agent can have a big impact on another agent's payoff without materially changing his own. Such mechanisms may have little hope of practical success if agents are inclined to manipulate each others' payoffs due to feelings of spite or kindness.

See Also

- ▶ Auctions (Experiments)
- ▶ Auctions (Theory)
- ▶ Contract Theory
- ▶ Hold-up Problem
- ▶ Incentive Compatibility
- ▶ Mechanism Design
- ▶ Revelation Principle

Bibliography

Abreu, D., and A. Sen. 1991. Virtual implementation in Nash equilibria. *Econometrica* 59: 997–1022.

Andreoni, J. 1995. Cooperation in public goods experiments: Kindness or confusion? *American Economic Review* 85: 891–904.

Armstrong, M. 1996. Multiproduct nonlinear pricing. *Econometrica* 64: 51–75.

Attiyeh, G., R. Franciosi, and R.M. Isaac. 2000. Experiments with the pivotal process for providing public goods. *Public Choice* 102: 95–114.

Baliga, S., and T. Sjöström. 1998. Decentralization and collusion. *Journal of Economic Theory* 83: 196–232.

Baliga, S., and T. Sjöström. 2006. Contracting with third parties. Working Paper No. 75, CSIO, Northwestern University.

Baliga, S., L. Corchón, and T. Sjöström. 1997. The theory of implementation when the planner is a player. *Journal of Economic Theory* 77: 15–33.

Bergemann, D., and S. Morris. 2005a. Robust mechanism design. *Econometrica* 73: 1771–1813.

Bergemann, D., and S. Morris. 2005b. Robust implementation: The role of large type spaces. Discussion Paper No. 1519, Cowles Foundation, Yale University.

Bergemann, D., and J. Välimäki. 2002. Information acquisition and efficient mechanism design. *Econometrica* 70: 1007–1034.

Cabrales, A., and G. Ponti. 2000. Implementation, elimination of weakly dominated strategies and evolutionary dynamics. *Review of Economic Dynamics* 3: 247–282.

Cabrales, A., G. Charness, and L. Corchón. 2003. An experiment on Nash implementation. *Journal of Economic Behavior and Organization* 51: 161–193.

Cason, T., T. Saijo, T. Sjöström, and T. Yamato. 2006. Secure implementation experiments: Do strategy-proof mechanisms really work? *Games and Economic Behavior* 57: 206–235.

Che, Y.K., and D. Hausch. 1999. Cooperative investments and the value of contracting. *American Economic Review* 89: 125–147.

Che, Y.K., and J. Kim. 2006. Robustly collusion-proof implementation. *Econometrica* 74: 1063–1107.

Chen, Y., and R. Gazzale. 2004. When does learning in games generate convergence to Nash equilibria? *American Economic Review* 94: 1505–1535.

Chen, Y., and F.F. Tang. 1998. Learning and incentive compatible mechanisms for public goods provision: An experimental study. *Journal of Political Economy* 106: 633–662.

Choi, J., and T. Kim. 1999. A nonparametric, efficient public decision mechanism: Undominated Bayesian Nash implementation. *Games and Economic Behavior* 27: 64–85.

Chung, K., and J. Ely. 2003. Implementation with near-complete information. *Econometrica* 71: 857–871.

Crémer, J. 1995. Arm's length relationships. *Quarterly Journal of Economics* 110: 275–295.

Crémer, J., and R. McLean. 1988. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica* 56: 1247–1257.

d'Aspremont, C., and L.A. Gérard-Varet. 1979. Incentives and incomplete information. *Journal of Public Economics* 11: 25–45.

Deb, R., P. Pattanaik, and L. Razzolini. 1997. Game forms, rights and the efficiency of social outcomes. *Journal of Economic Theory* 72: 74–95.

Dewatripont, M. 1989. Renegotiation and information revelation over time: The case of optimal labor contracts. *Quarterly Journal of Economics* 104: 589–619.

Dewatripont, M., and E. Maskin. 1995a. Credit and efficiency in centralized and decentralized economies. *Review of Economic Studies* 62: 541–555.

Dewatripont, M., and E. Maskin. 1995b. Contractual contingencies and renegotiation. *RAND Journal of Economics* 26: 704–719.

Duggan, J. 1997. Virtual Bayesian implementation. *Econometrica* 67: 1175–1199.

Edlin, A., and S. Reichelstein. 1996. Hold-ups, standard breach remedies and optimal investment. *American Economic Review* 86: 478–501.

Eliaz, K. 2002. Fault tolerant implementation. *Review of Economic Studies* 69: 589–610.

Eliaz, K., and R. Spiegler. 2006. Contracting with diversely naive agents. *Review of Economic Studies* 73: 689–714.

Esteban, S., and E. Miyagawa. 2005. Optimal menu of menus with self-control preferences. Unpublished paper, Penn State University.

Freixas, X., R. Guesnerie, and J. Tirole. 1985. Planning under incomplete information and the ratchet effect. *Review of Economic Studies* 52: 173–192.

Gaspart, F. 1995. Fair implementation in the cooperative production problem: Two properties of normal form mechanisms. Unpublished manuscript, FUNDP Namur.

Gul, F., and W. Pesendorfer. 2001. Temptation and self-control. *Econometrica* 69: 1403–1435.

Hart, O., and J. Moore. 1999. Foundations of incomplete contracts. *Review of Economic Studies* 66: 115–138.

Hurwicz, L., and S. Reiter. 2006. *Designing economic mechanisms*. New York: Cambridge University Press.

Jackson, M. 1991. Bayesian implementation. *Econometrica* 59: 461–477.

Jackson, M., T. Palfrey, and S. Srivastava. 1994. Undominated Nash implementation in bounded mechanisms. *Games and Economic Behavior* 6: 474–501.

Jehiel, P., and B. Moldovanu. 2001. Efficient design with interdependent valuations. *Econometrica* 69: 1237–1259.

Kawagoe, T., and T. Mori. 2001. Can the pivotal mechanism induce truthtelling? *Public Choice* 108: 331–354.

Krishna, V., and M. Perry. 1997. Efficient mechanism design. Unpublished manuscript, Penn State University.

Laffont, J.J., and D. Martimort. 1997. Collusion under asymmetric information. *Econometrica* 65: 875–911.

Maskin, E. 1992. Auctions and privatization. In *Privatization*, ed. H. Siebert. Tübingen: J.C.B. Mohr.

Maskin, E. 1999. Nash equilibrium and welfare optimality. *Review of Economic Studies* 66: 23–38.

Maskin, E., and J. Moore. 1999. Implementation and renegotiation. *Review of Economic Studies* 66: 39–56.

Maskin, E., and J. Tirole. 1999. Two remarks on the property rights literature. *Review of Economic Studies* 66: 139–150.

McAfee, P., and P. Reny. 1992. Correlated information and mechanism design. *Econometrica* 60: 395–421.

McLean, R., and A. Postlewaite. 2002. Informational size and incentive compatibility. *Econometrica* 70: 2421–2453.

Mezzetti, C. 2004. Mechanism design with interdependent valuations: Efficiency. *Econometrica* 72: 1617–1626.

Mezzetti, C. 2007. Mechanism design with interdependent valuations: Surplus extraction. *Economic Theory* 3: 473–499.

Mookherjee, D., and S. Reichelstein. 1992. Dominant strategy implementation of Bayesian incentive compatible allocation rules. *Journal of Economic Theory* 56: 378–399.

Mookherjee, D., and M. Tsumagari. 2004. The organization of supplier networks: Effects of delegation and intermediation. *Econometrica* 72: 1179–1219.

Nöldeke, G., and K. Schmidt. 1995. Option contracts and renegotiation: A solution to the hold–up problem. *RAND Journal of Economics* 26: 163–179.

Palfrey, T., and S. Srivastava. 1989a. Implementation with incomplete information in exchange economies. *Econometrica* 57: 115–134.

Palfrey, T., and S. Srivastava. 1989b. Mechanism design with incomplete information: A solution to the implementation problem. *Journal of Political Economy* 97: 668–691.

Palfrey, T., and S. Srivastava. 1991. Nash implementation using undominated strategies. *Econometrica* 59: 479–501.

Peleg, B., and E. Winter. 2002. Constitutional implementation. *Review of Economic Design* 7: 187–204.

Persico, N. 2004. Committee design with endogenous information. *Review of Economic Studies* 71: 165–191.

Postlewaite, A., and D. Schmeidler. 1986. Implementation in differential information economies. *Journal of Economic Theory* 39: 14–33.

Saijo, T. 1987. On constant Maskin monotonic social choice functions. *Journal of Economic Theory* 42: 382–386.

Saijo, T. 2003. Spiteful behavior in voluntary contribution mechanism experiments. In *Handbook of experimental economics results*, ed. C. Plott and V. Smith. Amsterdam: Elsevier Science.

Segal, I. 1999. Complexity and renegotiation: A foundation for incomplete contracts. *Review of Economic Studies* 66: 57–82.

Segal, I. 2004. The communication requirements of social choice rules and supporting budget sets. Working Paper No. 39, School of Social Science, Institute for Advanced Study, Princeton.

Segal, I., and M. Whinston. 2002. The Mirrlees approach to mechanism design with renegotiation. *Econometrica* 70: 1–46.
Sen, A. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78: 152–157.
Serrano, R., and R. Vohra. 2001. Some limitations of virtual Bayesian implementation. *Econometrica* 69: 785–792.
Sjöström, T. 1994. Implementation in undominated Nash equilibria without using integer games. *Games and Economic Behavior* 6: 502–511.
Sjöström, T. 1999. Undominated Nash implementation with collusion and renegotiation. *Games and Economic Behavior* 26: 337–352.
Smith, V. 1979. Incentive compatible experimental processes for the provision of public goods. In *Research in experimental economics*, vol. 1, ed. V. Smith. Greenwich: JAI Press.
Tirole, J. 1986. Hierarchies and bureaucracies. *Journal of Law, Economics, and Organization* 2: 181–214.
Williams, S. 1999. A characterization of efficient, Bayesian incentive compatible mechanisms. *Economic Theory* 14: 155–180.

Mechanism Design Experiments

Yan Chen and John O. Ledyard

Abstract

Mechanism design experiments bridge the gap between a theoretical mechanism and an actual economic process. In the domain of public goods, matching and combinatorial auctions and laboratory experiments identify features of mechanisms that lead to good performance when implemented among boundedly rational agents. These features include dynamic stability and security in public goods mechanisms, transparency in matching mechanisms, package bidding, simultaneity and iteration in combinatorial auctions.

Keywords

Bounded rationality; Combinatorial auction; Compensation mechanism; Convergence; Cooperation; Dynamic stability; English auction; Externalities; Matching; Mechanism design; Mechanism design experiments; Misrevelation; Public goods; Public goods production functions; Sealed bid auction; Social identity theory; Social loafing; Subgame perfection; Supermodularity; Value complementarities

JEL Classifications

C9

Mechanism design is the art of designing institutions that align individual incentives with overall social goals. Mechanism design theory was initiated by Hurwicz (1972) and is surveyed in Groves and Ledyard (1987). To bridge the gap between a theoretical mechanism and an actual economic process that solves fundamental social problems, it is important to observe and evaluate the performance of the mechanism in the context of actual decision problems faced by real people with real incentives. These situations can be created and carefully controlled in a laboratory. A mechanism design experiment takes a theoretical mechanism, recreates it in a simple environment in a laboratory with human subjects as economic agents, observes the behaviour of human subjects under the mechanism, and assesses its performance in relation to what it was created to do and to the theory upon which its creation rests. The laboratory serves as a wind tunnel for new mechanisms, providing evidence which one can use to eliminate fragile ones, and to identify the characteristics of successful ones.

When a mechanism is put to test in a laboratory, behavioural assumptions made in theory are seriously challenged. Theory assumes perfectly rational agents who can compute the equilibrium strategies via introspection. When a mechanism is implemented among boundedly rational agents, however, characteristics peripheral to theoretical implementations, such as transparency, complexity and dynamic stability, become important, or even central, to the success of a mechanism in a laboratory, and we suspect, ultimately in the real world. Mechanism design experiments cover several major domains, including public goods and externalities, matching, contract theory, auctions, market design and information markets. In what follows, we will review the experimental results of some of these topics.

Public Goods and Externalities

With the presence of public goods and externalities, competitive equilibria are not Pareto optimal. This is often referred to as market failure, since competitive markets on their own either result in underprovision of public goods (that is, the free-rider problem) or overprovision of negative externalities, such as pollution. To solve the free-rider problem in public goods economies, incentive-compatible mechanisms use innovative tax-subsidy schemes that utilize agents' own messages to achieve the Pareto optimal levels of public goods provision. A series of experiments test these mechanisms in the laboratory (see Chen 2008, for a comprehensive survey).

When preferences are quasi-linear, the Vickrey–Clarke–Groves (VCG) mechanism (Vickrey1961; Clarke 1971; Groves 1973) is strategy-proof, in the sense that reporting one's preferences truthfully is always a dominant strategy. It has also been shown that any strategy-proof mechanism selecting an efficient public decision at every profile must be of this type (Green and Laffont 1977). Two forms of the VCG mechanism have been tested in the field and laboratory by various groups of researchers. The pivot mechanism refers to the VCG mechanism when the public project choice is binary, while the cVCG mechanism refers to the VCG mechanism when the level of the public good is selected from a continuum. Under the pivot mechanism, misrevelation can be prevalent. Attiyeh et al. (2000) show that about ten per cent of the bids were truthfully revealing their values. Furthermore, there was no convergence tendency towards value revelation. In a follow-up study, Kawagoe and Mori (2001) show that more information about the payoff structure helps reduce the degree of misrevelation. More recently, Cason et al. (2006) provide a novel explanation for the problem of misrevelation in strategy-proof mechanisms. As Saijo et al. (2005) point out, the standard strategy-proofness concept in implementation theory has serious drawbacks, that is, almost all strategy-proof mechanisms have a continuum of Nash equilibria. They propose a new implementation concept, secure implementation, which requires the set of dominant strategy equilibria and the set of Nash equilibria to coincide. Cason et al. (2006) compare the performance of two strategy-proof mechanisms in the laboratory: the Pivot mechanism where implementation is not secure and truthful preference revelation is a weakly dominant strategy, and the cVCG mechanism with single-peaked preferences where implementation is secure. Results indicate that subjects play dominant strategies significantly more often in the secure cVCG mechanism (81 per cent) than in the non-secure Pivot mechanism (50 per cent). The importance of secure implementation in dominant strategy implementation is replicated in Healy (2006), where he compares five public goods mechanisms, voluntary contributions, proportional taxation, Groves_Ledyard, Walker and cVCG. The cVCG is found to be the most efficient of all mechanisms.

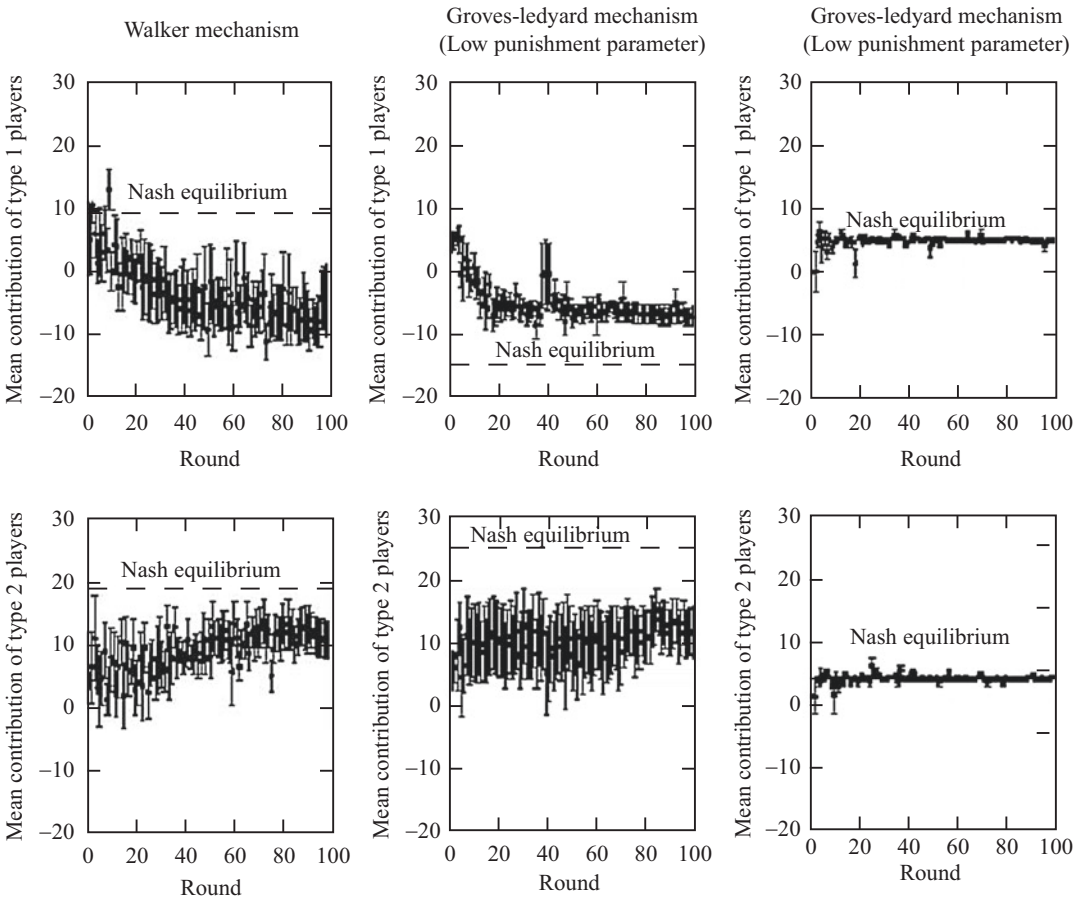
Although the VCG mechanism admits dominant strategies, the allocation is not fully Pareto-efficient. In fact, it is impossible to design a mechanism for making collective allocation decisions, which is informationally decentralized, non-manipulable and Pareto optimal. This impossibility has been demonstrated in the work of Hurwicz (1975), Green and Laffont (1977), Roberts (1979), Walker (1980) and Mailath and Postlewaite (1990) in the context of resource allocation with public goods.

Many 'next-best' mechanisms preserve Pareto optimality at the cost of non- manipulability, some of which preserve 'some degree' of non-manipulability. Some mechanisms have been discovered which have the property that Nash equilibria are Pareto optimal. These can be found in the work of Groves and Ledyard (1977), Hurwicz (1979), Walker (1981), Tian (1989), Kim (1993), Peleg (1996), Falkinger (1996) and Chen (2002). Other implementation concepts include perfect Nash equilibrium (Bagnoli and Lipman 1989), undominated Nash equilibrium (Jackson and Moulin 1992), subgame perfect equilibrium (Varian 1994), strong equilibrium (Corchon and Wilkie 1996), and the core (Kaneko 1977), and so forth. Apart from the above non-Bayesian mechanisms, Ledyard and Palfrey (1994) propose a class of Bayesian Nash mechanisms for public goods provision.

Experiments on Nash-efficient public goods mechanisms underscore the importance of dynamic stability, that is, whether a mechanism converges under various learning dynamics. Most of the experimental studies of Nash-efficient mechanisms focus on the Groves–Ledyard mechanism (Smith 1979a, b; Harstad and Marrese 1981, 1982; Mori 1989; Chen and Plott 1996; Arifovic and Ledyard 2006). Chen and Tang (1998) also compare the Walker mechanism with the Groves–Ledyard mechanism. Falkinger et al. (2000) study the Falkinger mechanism. Healy (2006) compares Nash-efficient mechanisms to cVCG and other benchmarks.

Among the series of experiments exploring dynamic stability, Chen and Plott (1996) first assessed the performance of the Groves–Ledyard mechanism under different punishment parameters. They found that by varying the punishment parameter the dynamics and stability changed dramatically. For a large enough parameter, the system converged very quickly to its stage game Nash equilibrium and remained stable; while under a small parameter, the system did not converge to its stage game Nash equilibrium. This finding was replicated by Chen and Tang (1998) with more independent sessions and a longer time series in an experiment designed to study the learning dynamics.

Figure 1 presents the time series data from Chen and Tang (1998) for two out of five types of players. Each graph presents the mean (the black dots) and standard deviation (the error bars) for each of the two different types averaged over seven independent sessions for each mechanism – the Walker mechanism, the Groves– Ledyard mechanism



Mechanism Design Experiments, Fig. 1 Mean contribution and standard deviation in Chen and Tang (1998)

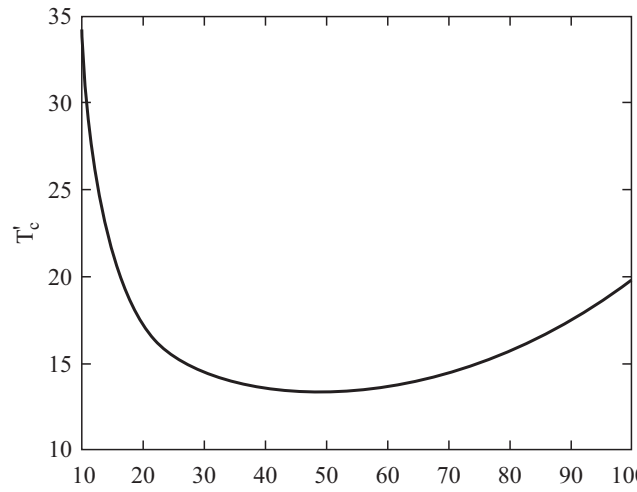
under a low punishment parameter (GL1), and the Groves–Ledyard mechanism under a high punishment parameter (GL100). From these graphs, it is apparent that GL100 converged very quickly to its stage game Nash equilibrium and remained stable, while the same mechanism did not converge under a low punishment parameter; the Walker mechanism did not converge to its stage game Nash equilibrium either.

Because of its good dynamic properties, GL100 had significantly better performance than GL1 and Walker, evaluated in terms of system efficiency, close to Pareto optimal level of public goods provision, fewer violations of individual rationality constraints and convergence to its stage game equilibrium.

These past experiments serendipitously studied supermodular mechanisms. Two recent studies systematically vary the parameters from below, close to, at and above the supermodularity threshold to assess the effects of supermodularity on learning dynamics.

Arifovic and Ledyard (2006) conduct computer simulations of an individual learning model in the context of a class of the Groves–Ledyard mechanisms. They vary the punishment parameter systematically, from extremely small to extremely high. They find that their model converges to Nash equilibrium for all values of the punishment parameter. However, the speed of convergence depends on the value of the parameter. As shown in Fig. 2, the speed of convergence is U-shaped: very low and very high values of the punishment parameter require long periods for convergence, while a range of intermediate values requires the minimum time. In fact, the optimal punishment parameter identified in the simulation is much lower than the supermudularity threshold. Predictions of the computation model are validated by experimental data with human subjects.

In a parallel research project on the role of supermodularity on convergence, Chen and Gazzale (2004) experimentally study the generalized version of the compensation mechanism (Varian 1994), which implements efficient allocations as subgame-perfect equilibria for economic environments involving externalities and public goods. The basic idea is that each player offers to compensate the other for the 'costs' incurred by making the efficient choice. They systematically vary the free parameter from below, close to, at and beyond the threshold of supermodularity to assess the effects of supermodularity on the performance of the mechanism. They have three main findings. First, in terms of proportion of equilibrium play and efficiency, they find that supermodular and 'near supermodular' mechanisms perform significantly better than those far below the threshold. This finding is consistent with previous experimental findings. Second, they find that from a little below the threshold to the threshold, the improvement in performance is statistically insignificant. This implies that the performance of 'near supermodular' mechanisms, such as the



Mechanism Design Experiments, Fig. 2 Convergence speed in Groves–Ledyard in Arifovic and Ledyard (2006)

Falkinger mechanism, ought to be comparable to supermodular mechanisms. Therefore, the mechanism designer need not be overly concerned with setting parameters that are firmly above the supermodular threshold – close is just as good. This enlarges the set of robustly stable mechanisms. The third finding concerns the selection of mechanisms within the class of supermodular mechanisms. Again, theory is silent on this issue. Chen and Gazzale find that within the class of supermodular mechanisms, increasing the parameter far beyond the threshold does not significantly improve the performance of the mechanism. Furthermore, increasing another free parameter, which is not related to whether or not the mechanism is supermodular, does improve convergence.

In contrast to the previous stream of work which identifies supermodularity as a robust sufficient condition for convergence, Healy (2006) develops a k-period average best response learning model and calibrates this new learning model on the data-set to study the learning dynamics. He shows that subject behaviour is well approximated by a model in which agents best respond to the average strategy choices over the last five periods under all mechanisms. Healy's work bridges the behavioural hypotheses that have existed separately in dominant strategy and Nash-efficient mechanism experiments.

In summary, experiments testing public goods mechanisms show that dominant strategy mechanisms should also be secure, while Nash implementation mechanisms should satisfy dynamic stability, if any mechanism is to be considered for application in the real world in a repeated interaction setting.

While experimental research demonstrates that incentive-compatible public goods mechanisms can be effective in inducing efficient levels of public goods provision, almost all the mechanisms rely on monetary transfers, which limit the scope of implementation of these mechanisms in the real world. In many interesting real world settings, such as open source software development and online communities, sizable contributions to public goods are made without the use of monetary incentives. We next review a related social psychology literature, which studies contribution to public goods without the use of monetary incentives.

Social Loafing

Analogous to free riding, social loafing refers to the phenomenon whereby individuals exert less effort on a collective task than they do on a comparable individual task. To determine conditions under which individuals do or do not engage in social loafing, social psychologists have developed and tested various theoretical accounts. Kauru and Williams (1993) present a review of this literature and develop a collective effort model, which integrates elements of expectancy value, social identity and self-validation theories, to explain social loafing. A metaanalysis of 78 studies shows that social loafing is robust across studies. Consistent with the prediction of the model, several variables are found to moderate social loafing. The following factors are of particular interests to a mechanism designer.

1. *Evaluation potential:* Harkins (1987) and others show that social loafing can be reduced or sometimes eliminated when a participant's contribution is identifiable and evaluable. In a related public goods experiment, Andreoni and Petrie (2004) find a substantial increase (59 per cent) in contribution to public goods compared to the baseline of a typical VCM experiment, when both the amount of individual contribution and the (photo) identification of donors are revealed.
2. *Task valence*: the collective effort model predicts that the individual tendency to engage in social loafing decreases as task valence (or perceived meaningfulness) increases.
3. *Group valence and group-level comparison standards:* Social identity theory (Tajfel and Turner 1986) suggests that 'individuals gain positive self-identity through the accomplishments of the groups to which they belong' (Kauru and Williams 1993, p. 686). Therefore, enhancing group cohesiveness or group identity might reduce or eliminate social loafing. In a closely related economics experiment, Eckel

and Grossman (2005) use induced group identity to study the effects of varying strength of identity on cooperative behaviour in a repeated public goods game. They find that while cooperation is unaffected by simple and artificial group identity, actions designed to enhance group identity contribute to higher levels of cooperation. This stream of research suggests that high degrees of group identification may limit individual shirking and free riding in environments with a public good.

4. *Expectation of co-worker performance influences individual effort*. This set of theories might be sensitive to individual valuations for the public good as well as the public goods production functions. The meta-analysis indicates that individuals loafed when they expected their co-workers to perform well, but did not loaf otherwise.
5. *Uniqueness of individual inputs:* individuals loafed when they believed that their inputs were redundant, but did not loaf when they believe that their individual inputs to the collective product were unique. In an interesting application, Beenen et al. (2004) conducted a field experiment in an online community called MovieLens. They found that users who were reminded of the uniqueness of their contributions rated significantly more movies than the control group.
6. *Task complexity:* individuals were more likely to loaf on simple tasks, but less likely on complex tasks. This finding might be related to increased interests when solving complex tasks.

Exploring non-monetary incentives to increase contribution to public goods is an important and promising direction for future research. Mathematical models of social psychology theories are likely to shed insights on the necessary and sufficient conditions for a reduction or even elimination of social loafing.

Matching

Matching theory has been credited as 'one of the outstanding successful stories of the theory of games' (Aumann 1992). It has been used to understand existing markets and to guide the design of new markets or allocation mechanisms in a variety of real world contexts. Matching experiments serve two purposes: to test new matching algorithms in the laboratory before implementing them in the real world, and to understanding how existing institutions evolved. We focus on one-sided matching experiments, and refer the reader to matching and market design for a summary of the two-sided matching experiments.

One-sided matching is the assignment of indivisible items to agents without a medium of exchange, such as money. Examples include the assignment of college students to dormitory rooms and public housing units, the assignment of offices and tasks to individuals, the assignment of students to public schools, the allocation of course seats to students (mostly in business schools and law schools), and timeshare exchange. The key mechanisms in this class of problems are the top trading cycles (TTC) mechanism (Shapley and Scarf 1974), the Gale–Shapley deferred acceptance mechanism (Gale and Shapley 1962), and variants of the serial dictatorship mechanism (Abdulkadiroglu and Sonmez 1998). Matching experiments explore several issues. For strategy-proof mechanisms, they explore the extent to which subjects recognize and use their dominant strategies without prompting. For mechanisms which are not strategy-proof, they explore the extent of preference manipulation and the resulting efficiency loss. As a result, they examine the robustness of theoretical efficiency comparisons when the mechanisms are implemented among boundedly rational subjects and across different environments.

For the class of house allocation problems, two mechanisms have been compared and tested in the laboratory. The random serial dictatorship with squatting rights (RSD) is used by many US universities for on-campus housing allocation, while the TTC mechanism is theoretically superior. Chen and Sonmez (2002) report the first experimental study of these two mechanisms. They find that TTC is significantly more efficient than RSD because it induces significantly higher participation rate of existing tenants.

Another application of one-sided matching is the time-share problem. Wang and Krishna (2006) study the top trading cycles chains and spacebank mechanism (TTCCS), and two status quo mechanisms in the time-share industry, that is, the deposit first mechanism and the request first mechanism, neither of which is efficient. In the experiment, the observed efficiency of TTCCS is significantly higher than that of the deposit first mechanism, which in turn, is more efficient than the request first mechanism. In fact, efficiency under TTCCS converged to 100 per cent quickly, while the other two mechanisms do not show any increase in efficiency over time.

More recently, the school choice problem has received much attention. We review two experimental studies. Chen and Sonmez (2006) present an experimental study of three school choice mechanisms. The Boston mechanism is influential in practice, while the Gale–Shapley and TTC mechanisms have superior theoretical properties. Consistent with theory, this study indicates a high preference manipulation rate under Boston. As a result, efficiency under Boston is significantly lower than that of the two competing mechanisms in the designed environment. However, contrary to theory, Gale–Shapley outperforms TTC and generates the highest efficiency. The main reason is that a much higher proportion of subjects did not realize that truth-telling was a dominant strategy under TTC, and thus manipulated their preferences and ended up worse off. While Chen and Sonmez (2006) examine these mechanisms under partial information, where an agent only knows his own preference ranking, and not those of other agents, a follow-up study by Pais and Pinter (2006), investigates the same three mechanisms under different information conditions, ranging from complete ignorance about the other participants' preferences and school priorities to complete information on all elements of the game. They show that information condition has a significant effect on the rate of truthful preference revelation. In particular, having no information results in a significantly higher proportion of truth-telling than under any treatment with additional information. Interestingly, there is no significant difference in the efficiency between partial and full information treatments. Unlike Chen and Sonmez (2006), in this experiment, TTC outperforms in terms of efficiency. Furthermore, TTC is also less sensitive to the amount of information in the environment.

Owing to their important applications in the real world, one-sided matching experiments provide insights on the actual manipulability of the matching mechanisms which are valuable in their real world implementations. Some issues, such as the role of information on the performance of the mechanisms, remain open questions.

Combinatorial Auctions

In many applications of mechanism design, theory is not yet up to the task of identifying the optimal design or even comparing alternative designs. One case in which this has been true is in the design of auctions to sell collections of heterogeneous items with value complementarities, which occur when the value of a combination of items can be higher than the sum of the values for separate items. Value complementarities arise naturally in many contexts, such as broadcast spectrum rights auctioned by the Federal Communications Commission, pollution emissions allowances for consecutive years bought and sold under the RECLAIM programme of the South Coast Air Quality Management District in Los Angeles, aircraft take-off and landing slots, logistics services, and advertising time slots. Because individuals may want to express bids for combinations of the items for sale, requiring up to 2^N bids per person when there are N items, these auctions have come to be known as combinatorial auctions.

As was discussed earlier under public goods mechanisms, theory has identified the VCG mechanism as the unique auction design that would implement an efficient allocation assuming bidders use dominant strategies. Theory has not yet identified the revenue-maximizing combinatorial auction, although Ledyard (2007) shows that it is not the VCG mechanism. Theory has also been of little use in comparing the expected revenue collection between different auction

designs. This has opened the way for many significantly different auction designs to be proposed, and sometimes even deployed, with little evidence to back up various claims of superiority.

To give some idea of the complexity of the problem we describe just some of the various design choices one can make. Should the auctions be run as a sealed bid or should some kind of iterative procedure be used? And, if the latter, should iteration be synchronous or asynchronous? What kinds of bids should be allowed? Proposals for allowable bids include only bids for a single item, bids for any package, and some which allow only a limited list of packages to be bid on. What stopping rule should be used? Proposals have included fixed stopping times, stop after an iteration in which revenue does not increase by more than *x* per cent, stop if demand is less than or equal to supply, and an imaginative but complex system of eligibility and activity rules created for the Federal Communications Commission (FCC) auctions. Should winners pay what they bid or something else? Alternatives to pay what you bid include VCG prices and second-best prices based on the dual variables to the programme that picks the provisional winners. What should bidders be told during the auction? Some designs provide information on all bids and provisional winners and the full identity of the bidders involved in them. Some designs provide minimal information such as only the winning bids without even the information as to who made them. The permutations and combinations are many. Because theory has not developed enough to sort out what is best, experiments have been used to provide some evidence.

The very first experimental analysis of a combinatoric auction can be found in Rassenti et al. (1982), where they compared a sealed bid auction (RSB) allowing package bids to a uniform price sealed bid auction (GIP), proposed by Grether et al. (1981), that did not allow package bids. Both designs included a double auction market for re-trading after the auction results were known. The RSB design yielded higher efficiencies than the GIP design. Banks et al. (1989) compared a continuous, asynchronous design (AUSM), a generalization of the English auction with package bidding, to a synchronous iterative design with myopic VCG pricing and found AUSM to yield higher efficiencies and revenues on average. Ledyard et al. (1997) compare the continuous AUSM to a synchronous iterative design (SMR) developed by Millgrom (2000) for the FCC auctions, which only allowed simultaneous single item bids. The testing found that ASUM yielded significantly higher efficiencies and revenues. Kwasnica et al. (2005) compare an iterative design (RAD) with package bidding and price feedback to both AUSM and SMR. RAD and SMR use the same stopping rule. Efficiencies observed with RAD and AUSM are similar and higher than those for SMR, but revenue is higher in SMR since many bidders lose money due to a phenomenon known as the exposure problem, which is identified in Bykowsky et al. (2000). If it is assumed that bidders default on bids on which they make losses and thus set the prices of such bids to zero, revenues are in fact higher under AUSM and RAD than under SMR. At the behest of the FCC, Banks et al. (2003) ran an experiment to compare an iterative, package bidding design (CRA) from Charles River Associates and Market Design (1998) with the FCC SMR auction format. They also found that the package bidding design provides more efficient allocations but less revenue, due to bidder losses in the SMR.

Parkes and Unger (2000) proposed an ascending price, generalized VCG auction (iBEA) that maintains nonlinear and non-anonymous prices on packages, and charges VCG prices to the winners. The design would theoretically produce efficient allocations as long as bidders bid in a straightforward manner. Straightforward bidding is myopic and non-strategic and involves bidding on packages that yield the locally highest payoff in utility. There is no evidence that actual bidders will behave this way. Chen and Takeuchi (2005) have experimentally tested iBEA against the VCG sealed bid auction and found that VCG was superior in both revenue generation and efficiency attained. Takeuchi et al. (2006) tested RAD against VCG and found that RAD generated higher efficiencies, especially in the earlier auctions. They were using experiments to test combinatoric auctions as a potential alternative to

scheduling processes in situations with valuation complementarities. In many cases current procedures request orderings from users and then employ a knapsack algorithm of some kind to choose good allocations without any concern for incentive compatibility. Takeuchi et al. (2006) find that both RAD and VCG yield higher efficiencies than the knapsack approach. Ledyard et al. (1996) found similar results when comparing a more vanilla combinatoric auction to an administrative approach. These findings suggest there are significant improvements in organization performance being overlooked by management.

Porter et al. (2003) proposed and tested a combinatorial clock (CC) auction. After bids are submitted, a simple algorithm determines the demand for each item by each bidder and for those items that have more than one bidder demanding more units than are available the clock price is raised. They test their design against the SMR and CRA. They do not report revenue but in their tests the CC design attained an almost perfect average efficiency of 99.9 per cent. CRA attained an average of 93 per cent, while SMR attained only 88 per cent. Brunner et al. (2006) have carried out a systematic comparison of SMR and three alternatives, CC, RAD and a new FCC design called SMRPB, which takes the basic RAD design and changes two things. SMRPB allows bidders to win at most one package and the pricing feedback rule includes some inertia that RAD does not. They find that in terms of efficiency RAD is better than CC which is equivalent to SMRPB which is better than SMR. In terms of revenue, they find CC is better than RAD which is better than SMRPB which is better than SMR.

Most of these papers compare only two or three auction designs at a time and the environments used as the basis for comparison is often different in different papers. Further, environments can often be chosen that favour one auction over another. To deal with this, many research teams stress test their results by looking at boundary environments' collections of payoff parameters that give each auction under examination its best or worst chance of yielding high revenue or efficiency. But it is still unusual for a research team to report on a comparative test of several auctions in which their own design ends up being outperformed by another. Nevertheless, there are some tentative conclusions one can draw from this research.

The easiest and most obvious conclusion is that allowing package bidding improves both efficiency and revenue. In all the studies listed, anything that limits bidders' ability to express the full extent of their willingness to pay for all packages does interfere with efficiency and revenue. Less obvious but also easy to see is that simultaneity and iteration are also good design features. Bidding in situations in which value complementarities exist can be difficult since bidders need to discover where their willingness to pay is more than others but also where they fit with others interests. Getting this right improves both efficiency and revenue. Iteration and relevant price feedback both help here. Stopping rules also matter. Although this is an area that could benefit from more research, it is clear that in many cases complicated stopping rules that allow auctions to proceed for very long periods of time provide little gain in revenue or efficiency.

Summary

Mechanism design experiments identify features of mechanisms that lead to good performance when they are implemented among real people. Experiments testing public goods mechanisms show that dominant strategy mechanisms should also be secure, while Nash-efficient mechanisms should satisfy dynamic stability if it is to be considered for application in the real world in a repeated interaction setting. For matching mechanisms, transparency of the dominant strategy leads to better performance in the laboratory. Lastly, in combinatorial auctions, package bidding, simultaneity and iteration are shown to be good design features. In addition to the three domains covered in this article, there has been a growing experimental literature on market design, information markets and contract theory. We do not cover them in this article, due to lack of robust empirical regularities. However, they are excellent areas in which to make a new contribution.

See Also

- ▶ Computing in Mechanism Design
- ▶ Matching and Market Design
- ▶ Mechanism Design
- ▶ Public Goods

Bibliography

Abdulkadiroglu, A., and T. Sönmez. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66: 689–701.

Andreoni, J., and R. Petrie. 2004. Public goods experiments without confidentiality: A glimpse into fundraising. *Journal of Public Economics* 88: 1605–1623.

Arifovic, J., and J. Ledyard. 2006. Computer testbeds and mechanism design: Application to the class of Groves-Ledyard mechanisms for provision of public goods. Caltech working paper. Pasadena.

Attiyeh, G., R. Franciosi, and M. Isaac. 2000. Experiments with the pivot process for providing public goods. *Public Choice* 102: 95–114.

Aumann, R. 1992. Foreword. In *Two-Sided matching: A study in game-theoretic modeling and analysis*, ed. E. Alvin, M.A. Roth, and O. Sotomayor. Cambridge: Cambridge University Press.

Bagnoli, M., and B. Lipman. 1989. Provision of public goods: Fully implementing the core through private contributions. *Review of Economic Studies* 56: 583–602.

Banks, J.S., J.O. Ledyard, and D.P. Porter. 1989. Allocating uncertain and unresponsive resources: An experimental approach. *Rand Journal of Economics* 20: 1–25.

Banks, J., M. Olson, D. Porter, S. Rassenti, and V. Smith. 2003. Theory, experiment and the federal communications commission spectrum auctions. *Journal of Economic Behavior and Organization* 51: 303–350.

Beenen, G., K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. Kraut. 2004. In *Proceedings of ACM computer supported cooperative work 2004,* conference on computer supported cooperative work. Chicago: ACM.

Brunner, C., J. Goeree, C. Holt, and J. Ledyard. 2006. Combinatorial auctioneering, Caltech working paper. Pasadena.

Bykowsky, M., R. Cull, and J. Ledyard. 2000. Mutually destructive bidding: The FCC auction design problem. *Journal of Regulatory Economics* 17: 205–228.

Cason, T., T. Saijo, T. Sjöström, and T. Yamato. 2006. Secure implementation experiments: Do strategy-proof mechanisms really work? *Games and Economic Behavior* 57: 206–235.

Charles River Associates Inc. and Market Design Inc. 1998. Report No. 1351–00.

Chen, Y. 2002. A family of supermodular Nash mechanisms implementing Lindahl allocations. *Economic Theory* 19: 773–790.

Chen, Y. 2008. Incentive-compatible mechanisms for pure public goods: A survey of experimental literature. In *The handbook of experimental economics results*, ed. D. Plott and V. Smith. Amsterdam: Elsevier.

Chen, Y., and R. Gazzale. 2004. Supermodularity and convergence: An experimental study of the compensation mechanism. *American Economic Review* 94: 1505–1535.

Chen, Y., and K. Takeuchi. 2005. Multi-object auctions with package bidding: An experimental comparison of Vickrey and iBEA. Working paper.

Chen, Y., and C.R. Plott. 1996. The Groves–Ledyard mechanism: An experimental study of institutional design. *Journal of Public Economics* 59: 335–364.

Chen, Y., and T. Sonmez. 2002. Improving efficiency of on-campus housing: An experimental study. *American Economic Review* 92: 1669–1686.

Chen, Y., and T. Sonmez. 2006. School choice: An experimental study. *Journal of Economic Theory* 127: 202–231.

Chen, Y., and F.-F. Tang. 1998. Learning and incentive compatible mechanisms for public goods provision: An experimental study. *Journal of Political Economy* 106: 633–662.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11: 17–33.

Corchon, L., and S. Wilkie. 1996. Double implementation of the ratio correspondence by a market mechanism. *Review of Economic Design* 2: 325–337.

Eckel, C., and P. Grossman. 2005. Managing diversity by creating team identity. *Journal of Economic Behavior & Organization* 58: 371–392.

Falkinger, J. 1996. Efficient private provision of public goods by rewarding deviations from average. *Journal of Public Economics* 62: 413–422.

Falkinger, J., E. Fehr, S. Gächter, and R. Winter-Ebmer. 2000. A simple mechanism for the efficient provision of public goods: Experimental evidence. *American Economic Review* 90: 247–264.

Gale, D., and L. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Green, J., and J.-J. Laffont. 1977. Characterization of satisfactory mechanisms for the revelation of the preferences for public goods. *Econometrica* 45: 427–438.

Grether, D., M. Isaac, and C. Plott. 1981. The allocation of landing rights by unanimity among competitiors. *American Economic Review* 71: 166–171.

Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.

Groves, T., and J. Ledyard. 1977. Optimal allocation of public goods: A solution to the 'free rider' problem. *Econometrica* 45: 783–809.

Groves, T., and J. Ledyard. 1987. Incentive compatibility since 1972. In *Essays in honor of Leonid Hurwicz*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Harkins, S.G. 1987. Social loafing and social facilitation. *Journal of Experimental Social Psychology* 23: 1–18.

Harstad, R.M., and M. Marrese. 1981. Implementation of mechanism by processes: Public good allocation experiments. *Journal of Economic Behavior & Organization* 2: 129–151.

Harstad, R.M., and M. Marrese. 1982. Behavioral explanations of efficient public good allocations. *Journal of Public Economics* 19: 367–383.

Healy, P.J. 2006. Learning dynamics for mechanism design: An experimental comparison of public goods mechanisms. *Journal of Economic Theory* 129: 114–149.

Hurwicz, L. 1972. On informationally decentralized systems. In *Decision and organization*, ed. C. McGuire and R. Radner. Amsterdam: North-Holland.

Hurwicz, L. 1975. On the existence of allocation systems whose manipulative Nash equilibria are Pareto-optimal. Paper presented at Third World Congress of the Econometric Society, Toronto.

Hurwicz, L. 1979. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *Review of Economic Studies* 46: 217–225.

Isaac, R., and D. James. 2000. Robustness of the incentive compatible combinatorial auction. *Experimental Economics* 3: 31–53.

Jackson, M., and H. Moulin. 1992. Implementing a public project and distributing its cost. *Journal of Economic Theory* 57: 125–140.

Kaneko, M. 1977. The ratio equilibria and the core of the voting game in a public goods economy. *Econometrica* 45: 1589–1594.

Kauru, S.J., and K.D. Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* 65: 681–706.

Kawagoe, T., and T. Mori. 2001. Can pivotal mechanism induce truth-telling? An experimental study. *Public Choice* 108: 331–354.

Kim, T. 1986. *On the nonexistence of a stable Nash mechanism implementing Lindahl allocations*. Mimeo: University of Minnesota.

Kim, T. 1993. A stable Nash mechanism implementing Lindahl allocations for quasi- linear environments. *Journal of Mathematical Economics* 22: 359–371.

Kwasnica, A.M., J.O. Ledyard, D. Porter, and C. DeMartini. 2005. A new and improved design for multi-object iterative auctions. *Management Science* 51: 419–434.

Ledyard, J. 2007. Optimal combinatoric auctions with single-minded bidders. *Proceedings of the 8th ACM conference on electronic commerce.* San Diego: ACM.

Ledyard, J., and T. Palfrey. 1994. Voting and lottery drafts as efficient public goods mechanisms. *Review of Economic Studies* 61: 327–355.

Ledyard, J., D. Porter, and C. Noussair. 1996. The allocation of a shared resource within an organization. *Economic Design* 2: 163–192.

Ledyard, J., D. Porter, and A. Rangel. 1997. Experiments testing multiobject allocation mechanisms. *Journal of Economics and Management Strategy* 6: 639–675.

Ledyard, J., M. Olson, D. Porter, J. Swanson, and D. Torma. 2002. The first use of a combined value auction for transportation services. *Interfaces* 32: 4–12.

Mailath, G., and A. Postlewaite. 1990. Asymmetric information bargaining problems with many agents. *Review of Economic Studies* 57: 351–367.

McAfee, P.R., and J. McMillan. 1996. Analyzing the airwaves auction. *Journal of Economic Perspectives* 10(1): 159–175.

McCabe, K., S. Rassenti, and V. Smith. 1989. Designing 'smart' computer assisted markets. *European Journal of Political Economy* 5: 259–283.

Milgrom, P., and J. Roberts. 1990. Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica* 58: 1255–1277.

Milgrom, P., and C. Shannon. 1994. Monotone comparative statics. *Econometrica* 62: 157–180.

Millgrom, P. 2000. Putting auction theory to work: The simultaneous ascending auction. *Journal of Political Economy* 108: 245–272.

Mori, T. 1989. Effectiveness of mechanisms for public goods provision: An experimental study. *Economic Studies* 40: 234–246.

Pais, J., and A. Pintér. 2006. School choice and information: An experimental study on matching mechanisms. Working paper, Institute for Economics and Business Administration (ISEG). Lisbon: Technical University.

Parkes, D., and L. Unger. 2000. Iterative combinatorial auctions: Theory and practice. In Proceedings of the 17th national conference on artificial intelligence (AAAI-00).

Peleg, B. 1996. Double implementation of the Lindahl equilibrium by a continuous mechanism. *Economic Design* 2: 311–324.

Porter, D.P. 1999. The effect of bid withdrawal in a multi-object auction. *Review of Economic Design* 4: 73–97.

Porter, D., S. Rassenti, A. Roopnarine, and V. Smith. 2003. Combinatorial auction design. *Proceedings of the National Academy of Sciences* 100: 11153–11157.

Rassenti, S., V. Smith, and R. Bulfin. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13: 402–417.

Roberts, J. 1979. Incentives and planning procedures for the provision of public goods. *Review of Economic Studies* 46: 283–292.

Saijo, T., T. Sjöström, and T. Yamato. 2005. Secure implementation. Working paper, no. 567–0047. Osaka: Institute of Social and Economic Research, Osaka University.

Shapley, L.S., and H. Scarf. 1974. On cores and indivisibilities. *Journal of Mathematical Economics* 1: 23–37.

Smith, V. 1979a. Incentive compatible experimental processes for the provision of public goods. In *Experimental economics*, vol. 1, ed. R. Smith and R. Smith. Greenwich: JAI Press.

Smith, V. 1979b. An experimental comparison of three public goods decision mechanisms. *Scandinavian Journal of Economics* 81: 198–251.

Tajfel, H., and J.C. Turner. 1986. The social identity theory of intergroup behaviour. In *Psychology of intergroup*

relations, ed. S. Worchel and W. Austin. Chicago: Nelson-Hall.
Takeuchi, K., J. Lin, Y. Chen, and T. Finholt. 2006. Shake it up baby: Scheduling with package auctions. Working paper. School of Information, University of Michigan.
Tian, G. 1989. Implementation of the Lindahl correspondence by a single-valued, feasible, and continuous mechanism. *Review of Economic Studies* 56: 613–621.
Varian, H. 1994. A solution to the problems of externalities when agents are well- informed. *American Economic Review* 84: 1278–1293.
Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16: 8–37.
Walker, M. 1980. On the impossibility of a dominant strategy mechanism to optimally decide public questions. *Econometrica* 48: 1521–1540.
Walker, M. 1981. A simple incentive compatible scheme for attaining Lindahl allocations. *Econometrica* 49: 65–71.
Wang, Y., and A. Krishna. 2006. Timeshare exchange mechanisms. *Management Science* 52: 1123–1137.

Medieval Guilds

Gary Richardson

Abstract

Guilds operated throughout Europe during the Middle Ages, and in many places into the early modern era. Merchant guilds were organizations of merchants involved in long-distance commerce and local wholesale trade, and may also have been retail sellers of commodities in their home cities and distant venues where they possessed rights to trade. Craft guilds were organized along lines of particular trades, their members typically owning and operating small family businesses. After the Black Death guilds grew rapidly in number, becoming the central economic and social institutions in medieval towns.

Keywords

Black Death; Collective liability; Craft guilds; Credit; Fertility; Free rider problem; Hanseatic League; Insurance; Medieval guilds; Merchant guilds; Monopoly; Reputation; Urbanization

JEL Classifications

N4

Guilds operated throughout Europe during the Middle Ages, and in many places, lasted into the early modern era. Guilds were groups of individuals with common goals whose activities, characteristics, and composition varied greatly across centuries, regions, and industries.

Guilds filled many niches in medieval economy and society. Typical taxonomies divide urban occupational guilds into two types: merchant and craft.

Merchant guilds were organizations of merchants who were involved in longdistance commerce and local wholesale trade, and may also have been retail sellers of commodities in their home cities and distant venues where they possessed rights to set up shop. The largest and most influential merchant guilds participated in international commerce and politics and established colonies in foreign cities. In many cases, they evolved into or became inextricably intertwined with the governments of their home towns.

Merchant guilds enforced contracts among members and between members and outsiders. Guilds policed members' behaviour because medieval commerce operated according to the community responsibility system. If a merchant from a particular town failed to fulfil his part of a bargain or pay his debts, all members of his guild could be held liable. When they were in a foreign port, their goods could be seized and sold to alleviate the bad debt. They would then return to their hometown, where they would seek compensation from the original defaulter.

Merchant guilds also protected members against predation by rulers. Rulers seeking revenue had an incentive to seize money and merchandise from foreign merchants. Guilds threatened to boycott the realms of rulers who did this, a practice known as *withernam* in medieval England. Since boycotts impoverished both kingdoms which depended on commerce and governments for whom tariffs were the principal source of revenue, the threat of retaliation deterred medieval potentates from excessive expropriations.

Craft guilds were organized along lines of particular trades. Members of these guilds typically owned and operated small businesses or family workshops. Craft guilds operated in many sectors of the economy. Guilds of victuallers bought agricultural commodities, converted them to consumables, and sold finished foodstuffs. Examples included bakers, brewers, and butchers. Guilds of manufacturers made durable goods and, when profitable, exported them from their towns to consumers in distant markets. Examples include makers of textiles, military equipment, and metalware. Guilds of a third type sold skills and services. Examples include clerks, teamsters, and entertainers.

These occupational organizations engaged in a wide array of economic activities. Some manipulated input and output markets to their own advantage. Others established reputations for quality, fostering the expansion of anonymous exchange and making everyone better off. Because of the underlying economic realities, victualling guilds tended towards the former. Manufacturing guilds tended towards the latter. Guilds of service providers fell somewhere in between. All three types of guilds managed labour markets, lowered wages, and advanced their own interests at their subordinates' expense. These undertakings had a common theme. Merchant and craft guilds acted to increase and stabilize members' incomes.

Non-occupational guilds also operated in medieval towns and cities. These organizations had both secular and religious functions. Historians refer to these organizations as social, religious, or parish guilds as well as fraternities and confraternities. The secular activities of these organizations included providing members with mutual insurance, extending credit to members in times of need, aiding members in courts of law, and helping the children of members afford apprenticeships and dowries.

The principal pious objective was the salvation of the soul and escape from Purgatory. Guilds served as mechanisms for organizing, managing, and financing members' collective quest for eternal salvation. Efforts centered on three types of tasks. The first were routine and participatory religious services such as prayers, processions, the singing of psalms, the illumination of holy symbols, and the distribution of alms to the poor. The second category consisted of actions performed on members' behalf after their deaths and for the benefit of their souls. Post-mortem services began with funerals and continued perpetually as guilds prayed (or hired priests to pray) for the salvation of the souls of all deceased members. The third category involved indoctrination and monitoring to maintain the piety of members.

Righteous living was important because members' fates were linked together. The more pious one's brethren, the more helpful their prayers, and the more quickly one escaped from purgatory. So, in hopes of minimizing purgatorial pain and maximizing eternal happiness, guilds beseeched members to restrain physical desires and forgo worldly pleasures.

Guilds also operated in villages and the countryside. Rural guilds performed the same tasks as social and religious guilds in towns and cities. Recent research on medieval England indicates that guilds operated in most, if not all, villages. Villages often possessed multiple guilds. Most rural residents belonged to a guild. Some may have joined more than one organization.

Guilds often spanned multiple dimensions of this taxonomy. Members of craft guilds participated in wholesale commerce. Members of merchant guilds opened retail shops. Social and religious guilds evolved into occupational associations. All merchant and craft guilds possessed religious and fraternal features.

In sum, guild members sought prosperity in this life and providence in the next. Members wanted high and stable incomes, quick passage through Purgatory, and eternity in heaven. Guilds helped them coordinate their collective efforts to attain these goals.

To attain their collective goals, guilds had to persuade members to contribute to the common good and deter free riding. Guilds that wished to develop respected reputations had to get all members to sell superior merchandise. Guilds that wished to lower the costs of labour had to get all masters to reduce wages. Guilds that wished to raise the prices of products had to get all masters to restrict output. Guilds whose members wished

to enter heaven had to get all members to live piously, abstaining both from the pleasures of the flesh and the material temptations of secular society.

To persuade members to cooperate and advance their common interests, guilds formed stable, self-enforcing associations that possessed structures for making and implementing collective decisions. A guild's members met periodically to elect officers, audit accounts, induct new members, debate policies, and amend ordinances. Officers administered a nexus of agreements among a guild's members. Details of these agreements varied greatly from guild to guild, but the issues addressed were similar in all cases. Members agreed to contribute certain resources or take certain actions that furthered the guild's occupational and spiritual endeavors.

Members who failed to fulfil their obligations faced punishments. Punishments varied across transgressions, guilds, time and place, but a pattern existed. First-time offenders were punished lightly, perhaps suffering public scolding and paying small monetary fines, and repeat offenders punished harshly. The ultimate threat was expulsion.

Within large guilds, a hierarchy existed. Masters were full members who usually owned their own workshops, retail outlets, or trading vessels. Masters employed journeymen, who were labourers who worked for wages on short-term contracts or a daily basis (hence the term journeyman, from *jour*, the French word for 'day'). Journeymen hoped to one day advance to the level of master. To do this, journeymen usually had to save enough money to open a workshop and pay for admittance, or, if they were lucky, receive a workshop through marriage or inheritance.

Masters also supervised apprentices, who were usually boys in their teens who worked for room, board and perhaps a small stipend in exchange for a vocational education. Both guilds and government regulated apprenticeships, usually to ensure that masters fulfilled their part of the apprenticeship agreement. Terms of apprenticeships varied, usually lasting from five to nine years.

Relationships between guilds and governments varied over centuries and around Europe. Guilds typically began as voluntary associations with little legal standing. Most guilds operated without formal recognition or authorization from the government. Successful occupational guilds aspired to attain recognition as a self- governing association with the right to possess property and other legal privileges. Merchant and craft guilds often purchased these rights from municipal and royal authorities.

The history of guilds stretches back to times with few written records. In the late Roman Empire, organizations resembling guilds existed in most towns and cities. These voluntary associations of artisans, known as *collegia*, were organized along trade lines. Members shared religious observances and fraternal dinners. Most of these organizations disappeared during the Dark Ages, when the Western Roman Empire disintegrated and urban life collapsed. In the Eastern Roman Empire, some *collegia* may have survived from late antiquity and evolved into medieval guilds, but it is unlikely that even the most resilient *collegia* survived in Western Europe.

In the centuries following the collapse of the Roman Empire, evidence indicates that guild-like associations operated in most towns and many rural areas. These organizations functioned as modern burial and benefit societies, whose objectives included prayers for the souls of deceased members, payments of *weregilds* in cases of justifiable homicide, and supporting members involved in legal disputes. These rural guilds were descendents of Germanic social organizations known as *gilda* which the Roman historian Tacitus referred to as *convivium*.

During the 11th through 13th centuries, considerable economic development occurred. The revival of long-distance trade coincided with the expansion of urban areas. Merchant guilds formed an institutional foundation for this commercial revolution. Merchant guilds sprung up in towns throughout Europe, and in many places rose to prominence in urban political structures. Merchant guilds' principal accomplishment was establishing the institutional foundations for long-distance commerce.

Merchant guilds first flourished in Italian cities in the 12th century. Craft guilds became ubiquitous

in Italy during the succeeding century. In northern Europe, merchant guilds rose to prominence a century later, when local merchant guilds in trading cities such as Lubeck and Bremen formed alliances with merchants throughout the Baltic region. The alliance system grew into the Hanseatic League which dominated trade around the Baltic and North Seas and in northern Germany.

As economic expansion continued in the 13th and 14 centuries, the influence of the Catholic Church grew, and the doctrine of Purgatory developed. The doctrine inspired the creation of countless religious guilds, since the doctrine provided members with strong incentives to want to belong to a group whose prayers would help one enter heaven and it provided guilds with mechanisms to induce members to exert effort on behalf of the organization.

The number of guilds grew rapidly after the Black Death, for several reasons. The decline in population raised per capita incomes, which encouraged the expansion of consumption and commerce, which in turn necessitated the formation of institutions to satisfy this demand. Repeated epidemics decreased family sizes, particularly in cities, where the typical adult had on average perhaps 1.5 surviving children, few surviving siblings, and only a small extended family, if any. Guilds replaced extended families in a form of fictive kinship. The decline in family size and impoverishment of the Church also forced individuals to rely on their guild more in times of trouble, since they no longer could rely on relatives and priests to sustain them through periods of crisis. All of these changes bound individuals more closely to guilds, discouraged free riding, and encouraged the expansion of collective institutions.

For nearly two centuries after the Black Death, guilds dominated life in medieval towns. Any town resident of consequence belonged to a guild. Most urban residents thought guild membership to be indispensable. Guilds dominated manufacturing, marketing, and commerce. Guilds dominated local politics and influenced national and international affairs. Guilds were the centre of social and spiritual life.

The heyday of guilds lasted into the 16th century. The Reformation weakened guilds. Afterwards, in Protestant nations the influence of guilds waned. Guilds often asked governments for assistance. Guilds requested monopolies on manufacturing and commerce and asked courts to force members to live up to their obligations. Guilds lingered where governments provided such assistance. Guilds faded where governments did not. Guilds retained strength in nations which remained Catholic until they were swept away by the reforms following the French Revolution and the Napoleonic Wars.

Bibliography

Basing, P. 1990. *Trades and crafts in medieval manuscripts*. London: British Library.

Cooper, R.C.H. 1985. *The archives of the city of London livery companies and related organizations*. London: Guildhall Library.

Davidson, C. 1996. *Technology, guilds, and early English drama.* Early drama, art, and music monograph series 23. Kalamazoo: Medieval Institute Publications/Western Michigan University.

Epstein, S.R. 1998. Craft guilds, apprenticeships, and technological change in preindustrial Europe. *Journal of Economic History* 58: 684–713.

Epstein, S. 1991. *Wage and labor guilds in medieval Europe*. Chapel Hill: University of North Carolina Press.

Gross, C. 1890. *The guild merchant: A contribution to British municipal history.* Oxford: Clarendon Press.

Gustafsson, B. 1987. The rise and economic behavior of medieval craft guilds: An economic-theoretical interpretation. *Scandinavian Journal of Economics* 35: 1–40.

Hanawalt, B. 1984. Keepers of the lights: Late medieval English parish guilds. *Journal of Medieval and Renaissance Studies* 14: 21–37.

Hatcher, J., and E. Miller. 1995. *Medieval England: Towns, commerce and crafts, 1086–1348*. London: Longman.

Hickson, C.R., and E.A. Thompson. 1991. A new theory of guilds and European economic development. *Explorations in Economic History* 28: 127–168.

Lopez, R. 1971. *The commercial revolution of the middle ages, 950–1350*. Englewood Cliffs: Prentice-Hall.

Mokyr, J. 1990. *The lever of riches: Technological creativity and economic progress*. Oxford: Oxford University Press.

Pirenne, H. 1952. *Medieval cities: Their origins and the revival of trade* (trans: Halsey, F.). Princeton: Princeton University Press.

Richardson, G. 2001. A tale of two theories: Monopolies and craft guilds in medieval England and modern imagination. *Journal of the History of Economic Thought* 23: 217–242.

Richardson, G. 2000. *Brand names before the industrial revolution*. Working paper, UC Irvine.
Richardson, G. 2004. Guilds, laws, and markets for manufactured merchandise in late-medieval England. *Explorations in Economic History* 41: 1–25.
Richardson, G. 2005a. Christianity and craft guilds in late medieval England: A rational choice analysis. *Rationality and Society* 17: 139–189.
Richardson, G. 2005b. The prudent village: Risk pooling institutions in medieval English agriculture. *Journal of Economic History* 65: 386–413.
Smith, T. 1870. *English Guilds*. London: N. Trübner & Co..
Swanson, H. 1983. *Building craftsmen in late medieval York*. York: University of York Press.
Thrupp, S. 1989. *The merchant class of medieval London 1300–1500*. Chicago: University of Chicago Press.
Unwin, G. 1904. *The guilds and companies of London*. London: Methuen & Company.
Ward, J. 1997. *Metropolitan communities: Trade guilds, identity, and change in early modern London*. Palo Alto: Stanford University Press.
Westlake, H.F. 1919. *The parish guilds of mediaeval England*. London: Society for Promotion of Christian Knowledge.

Meek, Ronald Lindley (1917–1978)

Andrew Skinner

Meek was born in Wellington, New Zealand, where he received his early education at both school and university. He went to Cambridge in 1946 to take a PhD under the supervision of Piero Sraffa (1949). Meek was appointed to a Lectureship in Glasgow University in 1948, during A.L. Macfie's tenure of the Adam Smith Chair.

Meek was translated to the Tyler Chair of Economics in Leicester University in 1963, where he did much to develop the Department. But it is as a lecturer that Ronald Meek is and was remembered by all those fortunate enough to have been taught by him. An admirable expositor, always prepared to an extent which included circulation of abstracts of the text, Meek's physical presence, allied to a stylish delivery, were admirably suited to the didactic tradition, especially in its Scottish form.

The high regard in which he was held by colleagues in Leicester is tangibly expressed in a volume of essays, edited by I. Bradley and M. Howard, entitled *Classical and Marxian Political Economy* (1982). This volume, and the attached bibliography, give some idea of Meek's output and range of interest. He wrote, for example, a series of articles in the field of Soviet Studies in the 1950s. These were followed in the 1960s by nine, highly technical, contributions to the study of the electricity industry. Meek's grasp of technique in this field of study may explain his later interest in quantitative methods; an interest which resulted in *Figuring Out Society* (1971). His last work would have been a book on matrix algebra.

In the 1960s Meek also found time to celebrate a life-long passion and a favourite place, in publishing what he always claimed to be his best-seller, *Hill Walking in Arran* (1963, 2nd edn, 1972).

But it is as an historian of economic thought that Meek will best be remembered; remembered for his contribution to the understanding of the classical period before Marx as well as for his essays on Marxian economics. Meek's position as a Marxist also helps to explain his *Studies in the Labour Theory of Value* (1956, 2nd edn, 1973) which owed 'its origin to a long correspondence which the author had in 1951 with Mrs. Robinson' regarding the validity of the labour theory of value (1956, p. 7). In this work, Meek sought to trace the historical development of the theory, before examining its restatement by Marx and its possible 're-application'.

Of the earlier writers, Meek's work on the French Oeconomists or Physiocrats is particularly noteworthy. The *Economics of Physiocracy* (1962) was followed by a variorum edition of Quesnay's *Tableau Oeconomique* (1972) and that in turn by translations of A.R.J. Turgot's historical essays (1973) which include the *Reflections on the Formation and Distribution of Riches*, written in 1766.

In translating these works, Ronald Meek did more than any other scholar to make them accessible to an English-speaking public. His extensive works of commentary also did more than any others to expose the purpose behind Quesnay's macroeconomic model of the 'circular flow' and threw a unique light on the still more sophisticated work of 'revisionists' such as Baudeau and

Turgot – with the latter producing a model of a capital-using system with distinct factors of production and categories of return. It is access of this kind which permits the modern scholar to form some estimate of the impact which such work must have had on Adam Smith when he visited Paris in 1766; a time when the intellectual output of the School had arguably reached its zenith (1962, pp. 31–3).

Meek's interest in Marx is also reflected in his identification of the historical and sociological (in addition to the economic) dimension of the work done by Quesnay and Turgot. This aspect of Meek's commentary also reflects his identification of what he called a 'Scottish Contribution to Marxist Sociology' (1954; 1967). The argument gave prominence to the 'four stages' theory of socio-economic development as it appeared especially in the work of Adam Smith and John Millar. Meek worked on this theme for the twenty years which preceded the publication of his *Social Science and the Ignoble Savage* (1976); the most complete statement of his position. The anthropological dimension reflects the content of his first published work (1943).

Meek's interest in the field is important of itself, but also for his appreciation of Adam Smith. Without suggesting that it is '*too* misleading' to imply that Smith was the author of a 'liberal' position (1977, p. 3) he felt it more important to note that:

> Smith, like Marx, was a whole man, who tried to combine a theory of history, a theory of ethics, and a theory of political economy into one great theoretical system. . ..There is no doubt that Marx can properly be said to be the heir of the basic ideas of the Scottish Historical School (1967, p. 50).

Such an appreciation of Smith helps to explain Meek's commitment to the planning of the Glasgow edition of the *Works and Correspondence*, following as it did on J.M. Lothian's discovery of new lecture notes in 1958. The same appreciation is evident in his meticulous preparation of the *Lectures on Jurisprudence* (Report dated 1762/63), for which he assumed the major responsibility.

Ronald Meek had a profound knowledge of Marx which informs and illuminates his works of commentary.

Selected Works

1943. *Maori problems today*. Wellington: Progressive Publishing House.

1956. *Studies in the labour theory of value*. London: Lawrence & Wishart. 2nd ed., 1973.

1962. *The economics of physiocracy: Essays and translations*. London: Allen & Unwin. This volume contains an introduction to the work of the school, and in Part 2 five essays, the first four of which appear in amended form. They are: 'Problems of the Tableau Economique' (1960); 'The Physiocratic Concept of Profit' (1959); 'Physiocracy and the Early Theories of Under-Consumption' (1951); 'Physiocracy and Classicism in Britain' (1951), and 'The Interpretation of Physiocracy' (1962).

1963. *Hill walking in Arran*. Isle of Arran Tourist Association. 2nd ed., 1972.

1967. *Economics and ideology and other essays: Studies in the development of economic thought*. London: Chapman & Hall. This Volume contains 12 essays which appear in amended form. They are: 'The Rehabilitation of Sir James Steuart' (1958); 'Adam Smith and the Classical Concept of Profit' (1954); 'The Scottish Contribution to Marxist Sociology' (1954); 'The Decline of Ricardian Economics in England' (1950); 'Thomas Joplin and the Theory of Interest' (1950–51); 'Karl Marx's Economic Method' (1959); 'The Falling Rate of Profit' (1960); 'Marx's Doctrine of Increasing Misery' (1962); 'Some Notes on the Transformation Problem' (1956); 'Mr. Sraffa's Rehabilitation of Classical Economics' (1961); 'The Place of Keynes in the History of Economic Thought' (1950–51); 'Economics and Ideology' (1957).

1971. *Figuring out society*. London: Collins.

1972. (With M. Kuczynski.) *Quesnay's 'Tableau Economique'*. London: Macmillan.

1973. *Turgot on progress, sociology and economics*. Cambridge: Cambridge University Press.

1976. *Social science and the ignoble savage*. Cambridge: Cambridge University Press.

1977. *Smith, Marx and after: Ten essays in the development of economic thought*. London:

Chapman & Hall. This vol. includes: 'Smith and Marx' (1977); 'Smith, Turgot and the Four Stages Theory' (1971); 'The Development of Adam Smith's Ideas on the Division of Labour' (with A.S. Skinner, 1973); 'New Light on Adam Smith's Glasgow Lectures on Jurisprudence' (1976); 'A Plain Person's Guide to the Transformation Problem' (1977); 'From Values to Prices: Was Marx's Journey Really Necessary?' (1976); 'The Historical Transformation Problem' (1976); 'Value in the History of Economic Thought' (1974); 'Marginalism and Marxism' (1972); 'The Rise and Fall of the Concept of the Economic Machine' (An Inaugural Lecture, 1965).

1978. (With D.D. Raphael and P.G. Stein.) *Adam Smith: Lectures on jurisprudence*. Oxford: Clarendon Press; Vol. V in the Glasgow edn of the *Works and Correspondence*.

A complete bibliography of R.L. Meek's writing will be found in *Classical and Marxian political economy: Essays in honour of Ronald L. Meek*, ed. I. Bradley and M. Howard, xi–xiv. London: Macmillan (1982).

Menger, Anton (1841–1906)

Andrea Ginzburg

Anton Menger, brother of the economist Carl, was a jurist and a socialist. Born in Maniow in Galicia on 21 September 1841, he died in Rome on 6 February 1906. After practising as a lawyer for some years, he became, in 1874, Professor of the Law of Civil Procedure at the University of Vienna, a post he was to hold until 1899. From 1886 onward he concentrated on an analysis of the legal aspects of socialism. In 1886 he published a study reconstructing the history of the claim of the worker to the whole produce of his labour. Though conducted explicitly from a legal standpoint, Menger's reconstruction – and, in particular, his rediscovery of English writers such as W. Godwin, C. Hall and W. Thompson, later to be described as (or associated with) 'Ricardian Socialists' – also attracted the interest of economists. This monograph was to prove highly influential both in German-speaking countries and beyond: in 1899 it was translated and published in England (with an introduction and bibliography by H.S. Foxwell) and in France (with an introduction by C. Andler). A strong opponent of Marxism, Menger's aim in his historical reconstruction was threefold. First and foremost, he asserted that 'the jurisprudential element' was, in fact, 'the real kernel of Socialism, in spite of the economic garb of which the modern socialists, more especially in Germany (Rodbertus, Marx, Lassalle) make so much' (English translation, p. 39). This underestimation of the economic element drew criticism from V. Pareto (1902–3, vol. 2, p. 86), amongst others. Menger also maintained that at the basis of socialist claims stood three 'economic rights': the right of the whole produce of one's labour, the right to subsistence, and the right to work. He held that the first of these was (or might be) incompatible with the second, and expressed his preference for the fulfilment of the right to subsistence by means of a system of public welfare. Foxwell, however, considers that 'the more novel side' and 'perhaps the occasion' of Menger's monograph lies in his determination to prove that Rodbertus and Marx 'borrowed their most important theories without any acknowledgement from many English and French theorists' (A. Menger, 1899, pp. xxv and cxv). In Menger's opinion, the author from whose writings Marx had 'borrowed' most freely was W. Thompson (Foxwell cites, in this context, J.F. Bray, while Andler cites Sismondi). Schumpeter observes that 'it is significant that this charge of plagiarism, though often repeated by economists was in the first instance raised by a writer who was not an economist himself' (Schumpeter 1954, p. 480). It may be added that the emphasis given by the jurist Menger to the problem of 'rights' probably contributed to the tendency of associating Ricardian socialists exclusively with this question, overshadowing in this way all those aspects of their thought which lie beyond it.

Selected Works

1886. *Das Recht auf der vollen Arbeitsertrag in geschichtlicher Darstellung*. Stuttgart.
1903. *Neue Staatslehre*. Jena.

Bibliography

Andler, C. 1899. Introduction to A. Menger, *Le droit au produit intégral du travail*. Paris.
Foxwell, H.S. 1899. Introduction to A. Menger. *The right to the whole produce of labour*. London. Reprinted, New York: Kelley, 1962.
Pareto, V. 1902–3. *Les systèmes socialistes*. In *Oeuvres Complètes*, vol. V, ed. G. Busino. Geneva: Librairie Droz, 1965.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.

Menger, Carl (1840–1921)

Karen I. Vaughn

Keywords

Austrian School; Bilateral competition; Böhm-Bawerk, E. von; Capital accumulation; Classical economics; Commodity money; Double coincidence of wants; Economic development; Economic good; Economic man; Exchange of equivalent values; Exchange theory; Fiat money; German Historical School; Hayek, F. A. von; Imputation; Isolated exchange; Labour theory of value; Limits of trade; Marginal productivity theory; Marginal utility analysis; Market intermediaries; Menger, C.; Methodenstreit; Methodological individualism; Mises, L. E. von; Money; Opportunity cost; Period of production; Precautionary balances; Price formation; Schmoller, G. von; Spontaneous order; Subjective theory of value; Time preference; Value; Wieser, F.F. von

JEL Classifications

B31

Carl Menger is known as one of the co-founders, along with W.S. Jevons and Leon Walras, of marginal utility analysis. As such, he can be counted as one of the originators of modern neoclassical economics. He is also recognized as the founder of the Austrian School of economics which developed a distinct tradition of economic thought over the century following his writing.

Menger was born in Neu-Sandez, Galizieu, a part of Austria that later became Poland. His family were mostly civil servants and army officers. Menger's father was a lawyer, and Carl studied law and political science first at the University of Vienna (1859–60) and then at Prague (1860–63). He took a doctorate at the University of Cracow and soon after, began a career in journalism. He worked in Lemberg and later Vienna where his main interests were in economic and fiscal problems of Austria. In 1871, Menger entered the Austrian civil service. However, 1871 was also the year in which his first book, *Grundsätze der Volkswirtschaftslehre*, later translated as *Principles of Economics*, was published. He presented this work for his habilitation for the faculty of law and political science at the University of Vienna. As a consequence, he became a 'privatdozent' and quit his position in the civil service. In 1873, he was appointed extraordinary professor and began his very long and very successful academic career. In 1876, Menger was appointed tutor to Crown Prince Rudolf of Austria and for two years accompanied him on travels through Germany, France, Switzerland and the British Isles. Upon his return, he resumed his teaching responsibilities, and he received a chair in political economy in 1879. He continued to teach until 1903 when, at the comparatively early age of 63, he retired to devote himself exclusively to completing the treatise he had begun with the *Principles*. He died within three days of his 81st birthday in 1921, his project still incomplete. He was survived by his one son, Karl.

The *Principles of Economics*

When Menger published the *Principles*, he was 31 years old and a journalist who had

recently been appointed to the prestigious 'Ministerratsprasidium' in the Austrian civil service. Several biographers report that during his years as a journalist Menger became interested in economics because he observed that current economic theories did not seem to explain current economic events. He therefore wanted to work out the laws of economics for himself. It is apparent from the internal textual evidence of the *Principles*, however, that Menger must have had a more than cursory interest in the subject of economics during his years as a student. He must have read deeply and widely in the history of economics, since his first major work cites a wide range of earlier thinkers on economic problems including Aristotle, the medieval scholastics, Turgot, Smith, Ricardo, the German historicists and the contemporary socialists. Menger's knowledge of the history of economic thought is also evidenced by the outstanding library he accumulated during his lifetime, and by the fact that most of the major works in economic thought bear the marks of his close study.

Menger's clear purpose was to show how his theory of value could solve satisfactorily and in a unified manner all the problems of economic theory posed by earlier thinkers. The major target of his work was the labour theory of value which he believed was not only incorrect as an explanation of value and prices, but also failed to provide a unified explanation for factor prices on its own terms. However, Menger also took as his task to explain away the paradox of value, the erroneous view of Aristotle that exchange was an exchange of equivalent values, the mistaken view that capital as such was productive and the notion that money had to be explained according to different principles from other goods. In fact, every chapter of the *Principles* contains a refutation of some earlier doctrine or other that required a correct theory of value to elucidate.

Menger was writing partly against the background of classical economics. He was also, however, writing to an audience of German scholars who, in their rejection of classical economics were also rejecting the whole notion that one could develop a scientific theory of economic phenomena. Nothing, however, could be further from Menger's approach. Part of his aim, then, was to explain to the German historical economists that scientific economic theory was possible and compatible with empirical reality. To that end, Menger dedicated the book 'with respectful esteem to D. Wilhelm Roscher', a major figure in the older German Historical School.

To Menger, the central unifying principle of economics was the phenomenon of value. One had to explain the source of value before any of its particular manifestations could be understood. However, to develop adequately a theory of value, Menger had to prepare the ground upon which the theory rests. For Menger, that meant spending the first two chapters of his book (62 pages and over one-fourth the main text) on an exhaustive discussion of the meaning of a good and of an economic good in particular. While to the modern reader this might seem excessively thorough, Menger, the innovator, wished to take nothing for granted in establishing the firm basis of his theory. One had to move from the notion of useful things to the notion of a good to the concept of an economic good before one could understand the real meaning of economic value. Since all of economic theory hangs on this concept, he must have believed it imperative to be sure the reader understands each step of the argument.

Right from the beginning we see Menger's distinctive approach to economic theory. 'All things are subject to the law of cause and effect' (*Principles*, p. 51). Economic theory is an exercise in discovering and explaining the causal relationship between things and human values. He thus begins by pointing out that there are many useful things in the world, but for a useful thing to have 'goods-character', men must (*a*) recognize a causal connection between the good and its ability to satisfy a need and (*b*) have the power to make use of the thing for need satisfaction (Goods, Menger points out, can be concrete things or they can be intangible relationships such as firms, copyrights or good will, an observation that is distinctly modern). This is a pattern repeated again and again in Menger's writing: men must have knowledge and power. Economic life is built around gaining knowledge and power; knowledge of causal relationship between things and satisfaction, knowledge of technical

production relationships, knowledge of trading opportunities, knowledge of 'economic' prices, knowledge of the qualities of goods, and the power to make the best use of man's knowledge.

Knowledge of causal connections among goods permits men to rank goods in accordance with their relationship to want satisfaction. Goods that have the ability to satisfy needs directly (consumer goods), Menger called 'first order goods'. Goods that can only indirectly satisfy needs by being transformed with complementary goods into first order goods, Menger called 'goods of a higher order' (inputs). Furthermore, higher order goods are not valued in themselves, but derive their 'goods character' from first order goods, an observation that will later allow Menger to develop his refutation of the labour theory of value.

Having established the concept of a good in general in Chapter 1, Menger goes on in Chapter 2 to explain the concept of an economic good. Menger's definition of an economic good is completely familiar to modern readers; the way in which Menger develops his argument is not. Menger sees men's strictly economizing activities as taking place within the context of an overall plan through time. He argues that men must estimate both their needs for various goods and the quantities of goods that will be available for fulfilling their consumption plans for specific periods of time. Their estimated needs, he calls their 'requirements' (*bedarfs*) a concept for which we have no modern equivalent although Stigler (1941, p. 140) has argued that requirements are the quantities of goods sufficient to make marginal utility go to zero (all the economic goods men could rationally consume). An economic good, then, is one where available quantities fall short of men's requirements.

The notion of requirements is important to Menger's argument because it allows him to discuss how men get information about requirements and quantities of goods and how they plan for their consumption in the face of uncertainty. It is obvious in this discussion that Menger does not hypothesize given utility functions which are maximized subject to fixed constraints. While he eventually gets to a verbal explanation of economizing behaviour that is consistent with the standard model, to him the interesting questions involve how men go about estimating their requirements over time and how they plan to satisfy them. Their planning activities require not only that they estimate future needs based on present tastes and preferences, but that they take into account the fact that their needs may change in unexpected ways. Further, their planning activity also encompasses plans to change the quantities of goods available. Hence, the plan is one of production as well as consumption. Only after Menger establishes the importance of human planning through time does he go on to discuss economizing behaviour in the modern sense of maximizing satisfaction within the known resource constraints.

Economizing to Menger, then, is a two-step process that involves first formulating a general plan for meeting one's requirements by assessing probable needs given an uncertain future and gathering information about the probable availability of goods, and then actually economizing based on the actual needs and quantities available at a moment in time.

Menger's discussion of an economic good is rich with associated insights. In this chapter, he gives an account of how non-economic goods become economic (through growth of population, growth of human needs and advances in knowledge as civilization progresses), a description of public goods (goods that are economic goods in general but are provided in such a way that people treat them as non-economic goods), an account of the origin and function of private property (to protect economic goods owned by the haves from the predation of the have-nots). Property is the 'only practically possible solution of the problem that is, in the nature of things, imposed upon us by the disparity between requirements for, and available quantities of, all economic goods' (*Principles*, p. 97), and a discussion of the economic implications of differing qualities of goods. He devotes part of his discussion of economic goods to discussing the nature of individual wealth – the entire sum of economic goods at an individual's command – and of national wealth – a slippery concept that can only be accurately described as 'a complex of wealths linked together by intercourse and trade' (*Principles*, p. 112).

Finally, after this detailed groundwork, Menger gets to his theory of value in Chapter 3. Menger has been called a member of the 'psychological school' because of his thoroughly subjective notion of value. However, it is not a Jevons-like utilitarian subjectivism. Goods are valued not because they provide various quantities of utils to individuals, but because they serve various uses that have different levels of importance to individuals. The difference may seem small to the reader, but it makes for subtle but important differences in understanding the valuation process. 'Value is . . . the importance that individual goods or quantities of goods attain for us because we are conscious of being dependent on command of them for the satisfaction of our needs' (*Principles*, p. 115). Value is a judgement men make about the importance of goods; it adheres in concrete units of goods and not in abstract utility. The problem of a theory of value is to explain the differences in value among different goods.

Menger develops his theory of value in two stages. First he shows, with the use of a numerical table, how the importance men attach to the acquisition of additional units of a good that satisfies a particular need declines as more of the good is acquired, and by comparing the declining satisfactions associated with the acquisition of increasing amounts of various goods, he explains why a man might satisfy some of his desire for tobacco, for example, before he has completely satisfied his desire for food. In fact, Menger's tables are vivid examples of Gossen's first and second laws. Menger's use of numbers may give the impression that he is explaining utility as a cardinally measurable quantity. However, the impression is immediately dispelled when he points out that his chart is merely illustrative of a general psychological principle and is not meant to be taken literally. Furthermore, his chart, he explains, describes only a special case of valuation – the case where a single good serves for a single satisfaction. The more important case – where a single good has multiple uses – is more complex and requires more discussion. Interestingly, it isonly in the context of the following more complex case that he states clearly his principle of diminishing marginal valuation.

When a single good, such as sacks of grain or pails of water, can serve many different uses, the first units will be used to serve the most important uses for the good while successive units of the good will be put to less and less important uses. Menger concludes, then, that the value of any one sack of grain is equal to the satisfaction associated with the least important use that would go unsatisfied if one sack of grain is removed, a statement of diminishing marginal utility that is completely free of mathematical metaphor.

Menger drew two immediate implications from his value theory: (1) the diamonds–water paradox was easily solved because given their respective quantities, the marginal unit of water in most cases served no use while the marginal unit of scarce diamonds had very important desires to satisfy, and (2) the labour theory of value was obviously incorrect.

> The determining factor in the value of a good, then, is neither the quantity of labor or other goods necessary for its production not the quantity necessary for its reproduction, but rather the magnitude of importance of those satisfactions with respect to which we are conscious of being dependent on command of the good. This principle of value determination is universally valid, and no exception to it can be found in human economy. (*Principles*, p. 145)

This leads Menger to one of the most important theoretical implications of his theory – that the value of goods of a higher order depends on the prospective value of corresponding goods of lower order. In fact, the value of an input is equal to the satisfaction that would be forgone if the input were not available for use. Note that this is not so much a marginal productivity theory of factor value as it is a 'marginal utility product' theory completely consistent with his subjective theory of value.

Despite his comments on the value of goods of a higher order, Menger did not develop a theory of production in the modern sense. He did observe, however, that all production takes place in time, and that the higher the order of goods employed, the more distant in time will be the final satisfaction obtained. The only way men can increase output is 'to lengthen the period of time over which their provident activity is to extend in the

same degree that they progress to goods of higher order' (*Principles*, p. 153). This suggestion was the basis upon which Böhm-Bawerk constructed his theory of the period of production that led to so much controversy by the end of the 19th century. Menger also points out that the limit to economic progress is the degree to which men value the same satisfaction more highly in the present rather than the future. Later called 'time-preference' by Austrian economists, Menger believed it was a consequence of men's continuous and finite life span. Without time preference, one would have to expect infinite capital accumulation. Notice that time preference is an explanation for why there is a limit to capital accumulation rather than an explanation for why capital is accumulated at all.

Menger is best known for his theory of value and its implication for goods of a higher order. His theory of exchange and price is neither so well-known nor so highly regarded. This is a pity since the chapters following the theory of value are equally rich with economic insights and deserve close attention by modern readers. Predictably, Menger's theory of exchange is derived from his theory of value. His starting point is Adam Smith's statement that men are possessed of a 'propensity to truck, barter and exchange one good for another', a statement Menger finds objectionable since it provides no explanation for the particular kinds of trade men make or for the limits of their trading activity. Men do not trade because of a propensity to do so, but because of a rational desire to improve their well-being. Men seek out trade opportunities in order to exchange something less valuable for something more valuable and hence trade is productive of value for both trading partners. The problem for the economist, then, is to determine the limit of trade, limits that will be reached when neither party any longer stands to gain.

While Menger's theory of trade is fairly standard, less standard is his very modern discussion of the importance of transactions costs in limiting trade. These 'economic sacrifices of exchange' (*Principles*, p. 189) arise because men and their possessions are separate in space and time and must be brought together for trade to take place. Sometimes these economic sacrifices are so great that a potentially productive trade does not take place at all. It is the role of market intermediaries (including entrepreneurs) to reduce the economic sacrifices of trade through improved knowledge and improved market organizations. Entrepreneurs bring together potential traders, and the source of the intermediary's income is the gain in satisfaction permitted by his activities. The idea of transactions costs and the role of market 'intermediaries' in reducing transactions costs was rediscovered in the 1950s.

Menger's theory of exchange leads him to develop his theory of price. This chapter eventually arrives at propositions that are now standard in price theory, but it does so in a peculiar way. Menger states in the very beginning of the chapter that contrary to the beliefs of some earlier thinkers, price is not the fundamental feature of exchange. While price is directly observable, it is derivative of the real fundamental feature of exchange: the utility gain from trade. Price is merely a 'symptom of an economic equilibrium between the economies of individuals' (*Principles*, p. 191). There should be no misunderstanding then about exchange involving an exchange of equivalent values. If such were the case, men would be willing to reverse their trades since there would be no gain or loss involved. But we do not observe such 'reversible' trades in the real world because trades are not of equivalent values but of subjective values that differ for each party to the trade. Price theory, then, is not a theory of establishing equivalents for exchange but rather a theory that seeks to explain why men give specific quantities of goods for specific quantities of other goods.

Menger approaches this problem in a way that was to become common in neoclassical economics – according to the number of traders in the market. However, instead of taking the case of many buyers and sellers as the norm and examining various monopolies as deviations, he begins with the simplest case of two party exchange ('isolated exchange') and progresses through various monopoly models finally to reach the case of 'bilateral competition'. The reason for this progression is not simply analytic simplicity; he believed that this was the way trade actually

developed in history with monopolies giving way to more and more competitive conditions, and he gives several historical examples to support his case.

Under isolated exchange, price will fall within a range set by the marginal utilities of the two traders. The actual price is indeterminate from the point of view of theory, but in most cases, Menger argued, neither party will have any special bargaining power and they will agree to a price that gives them a more or less equal utility gain.

From there Menger progresses to the case where a monopolist provides a single good to several competing buyers. In this case, the limits within which price will fall are narrowed by the intensity of demand of the most eager buyer and the one next most eager to acquire the good.

The case of monopoly provision of several units of a good to competing buyers is even more interesting. There, assuming a uniform price is established:

> price formation takes place between the limits that are set by the equivalent of one unit of the monopolized good to the individual least eager and least able to compete who still participates in the exchange and the equivalent of one unit of the monopolized good to the individual most eager and best able to compete of the competitors who are economically excluded from the exchange. (*Principles*, p. 207).

One important implication is that the larger the quantity offered for sale by the monopolist, the 'lower in terms of purchasing power and eagerness to trade will he have to descend among the classes of competitors for the monopolized good in order to sell the whole quantity, and hence the lower also will be the price of one unit of the monopolized good' (*Principles*, p. 207). In this way, Menger established the inverse price–quantity relationship that had been assumed by economists prior to the introduction of marginalism into economic thought.

What is interesting in Menger's approach is that he emphasizes that the process of price formation is the same regardless of the market conditions. Monopolists are subject to the limits placed on their actions by the utilities of the buyers just as competitors are so limited. What does vary according to market conditions are the policies of sellers. Monopolists may well follow a policy of restricting supply in order to sell few units at higher prices, or they may follow a policy of selling different units at different prices depending on the buyers. Competitors in supply of a product, however, will never find those policies to their advantage and hence under bilateral competition, one would expect prices to be lower and quantities supplied to be greater.

There is some debate as to whether Menger was offering an equilibrium theory of price determination in the *Principles* (Streissler 1972; Jaffé 1976). Certainly, his method of reasoning implies some underlying equilibrium price within any given market, and he even states that from time to time equilibrium prices will be observed. Equilibrium prices are 'economic' prices in that transactions at these prices are the result of economizing behaviour where no one could have been better off at another price. Further, he describes prices that reflect the full 'economic situation', a phrase that seems to indicate a more widespread economic equilibrium. However, it is also true that Menger did not describe economies settling down to a strict general equilibrium in the manner of Walras. Indeed, given the barriers to strict economic behaviour, especially barriers of incomplete knowledge, that are inherent in real life, Menger would find a Walrasian general equilibrium in principle unattainable. Men did the best they could, and with economic progress their best got better, but the very conception of a Walrasian general equilibrium is foreign to Menger's method of reasoning. This will become clearer below when Menger's methodology is discussed.

The next two chapters, 'Use Value and Exchange Value' and 'The Theory of the Commodity', while containing several interesting discussions about market organization, are really prelude to the very important last chapter on 'The Theory of Money'. In the 'Commodity' chapter, Menger defines a commodity as a good intended for sale and then discusses the varying degrees of saleability of commodities based on their characteristics and market organization. The point he isleading to is to define money as the most marketable of all commodities, his starting point for the last chapter.

Menger does not develop a theory of the value of money in the Principles. While he does stress the importance of holding precautionary balances, to Menger the most important questions are how does money come to exist and what functions does it serve. These are questions he addresses both in his *Principles*, in the later work on methodology and in his two articles on money written in 1892. From a modern perspective, there are two particularly interesting features of Menger's discussion that should be noted. First, Menger's account of the origin of money is developed in a way reminiscent of the reasoning of the Scottish Enlightenment and Adam Smith's 'invisible hand' in particular (although it is doubtful that the writers of the Scottish Enlightenment were the direct sources for his reasoning. In fact, at one point, he criticizes Adam Smith for a too mechanistic and rationalistic view of economic and social institutions! [*Investigations*, p. 177]). Money, according to Menger, arises out of the self-interested actions of individuals aimed at attaining their own ends through trade, but not specifically aimed at developing a money commodity as such. Second, because money arises as an unintended by-product of human action, it is not a creation of government.

The process Menger describes for the origin of money is a straightforward extension of his theory of economizing behaviour through trade. Following Aristotle, Menger points out the difficulties men face under barter in finding trading partners, (the problem of the 'double coincidence of wants'). Rational men soon come to realize that goods have different degrees of marketability. A cow, for instance, is far more marketable than custom made shoes. Hence, men learn that if they exchange their less marketable goods for goods that may not directly satisfy their needs but that are more marketable, they will be more successful at bartering for what they really need. Eventually, Menger reasons, some one commodity will emerge as the most marketable commodity and men will be willing always to accept it in exchange for other goods because they know they will have no trouble trading it for what they really want. This most marketable commodity then becomes money. While specific money commodities have differed from one society to another, in the most developed countries, precious metals become the money commodity because of their suitable characteristics: their portability, divisibility, scarcity, and so on.

Obviously, in such a theory money cannot be a creation of government because it is a naturally evolved social institution. Government can enhance the acceptability of a money commodity by declaring it legal tender, but government cannot create money. In this way, Menger's theory is meant to solve several long-standing controversies in the theory of money. The nominalist–realist debate is resolved by acknowledging that the value of money commodity is equal to the value of the money (except for small coins where it would be uneconomic to spend the resources to make full-weight coins) but by also pointing out that the actual commodity can be anything consistent with the accepted standards and level of development of the community. The commodity–fiat debate is resolved by showing a role for government in enhancing the acceptability of money even though it originates first through a natural process of human choices.

The last chapter is not the only place in which Menger discusses the origin of an economic phenomenon. All through the *Principles*, Menger is interested in establishing the origin and meaning of phenomena where the meaning is often elucidated through a description of their evolution through time. Erich Streissler (1972, p. 430) has gone so far as to credit Menger with presenting foremost a theory of economic development in the *Principles*. There is much to recommend that position. One of Menger's main themes is that economic development is a process of increasing knowledge and the consequent improvement in the variety and quality of goods available. Economic development is characterized by better communication among traders, more complex trading institutions, more and better commodities, and a greater ability of men to establish 'economic' prices.

We can perhaps understand Menger's vision better if we remember how he thought of the human predicament. Man in his original state is ignorant of his environment and uncertain about

his (finite) future. He must plan for the satisfaction of his wants in this difficult world, and his primary aid is his ability to learn. The progress of civilization is nothing so much as a process of reducing ignorance and developing institutions that make dealing with the uncertain future more manageable. Smith emphasized the division of labour and capital accumulation as the causes of the wealth of nations. Menger emphasized the priority of improved knowledge to the improvement of wealth. Indeed, progress is evidenced by 'the increasing understanding of the causal connections between things and human welfare' (*Principles*, p. 74).

Methodology

While Menger's *Principles* was well received and eventually became very influential in his native Austria, his theories came in for criticism – or, more to the point, apathy and neglect – in the one audience Menger had most hoped to convince, the German Historical School. While the older members of the historical school, Knies, Roscher and Hildebrand, understood classical economic theory and wanted to overcome its shortcomings with detailed historical investigations which would have the purpose of allowing them to infer them empirical regularities in economic events, the younger Germans, led by Gustav Schmoller, rejected the theory entirely. They believed there could be no such thing as scientific economic theory, and they insisted on viewing an economy as an organic whole at one with politics, law and custom. Menger's new theory, then, was considered not only incorrect, but useless. To Menger, who was convinced that he had discovered the key to unlocking the mysteries of all economic phenomena, such cavalier dismissal must have been particularly galling.

Having failed to make headway with his new theory in Germany on what appeared to be methodological grounds, Menger began work in 1875 on his second book, *Untersuchungen über die Methode der Socialwissenschaften und der politischen Oekonomie insbesondere* (*Investigations into the Method of the Social Sciences with special reference to Economics*). This book, essentially a defence of economic theory and an account of its relationship to historical methods, was published in 1883. Menger's ambition was this time to attract the attention of German academics. This time he succeeded, but unfortunately, the attention he attracted was negative. Gustav Schmoller's review of the Investigations was particularly unsympathetic and incited Menger to respond with an impassioned pamphlet entitled *The Errors of Historicism* in 1884. In this pamphlet, Menger dropped all attempts at cordial conciliation and, in Hayek's words, 'ruthlessly demolished Schmoller's position' (Hayek 1981, p. 24). If so, Schmoller never discovered the demolition since he returned the book to Menger unread and wrote a final scathing attack on Menger in his journal.

This exchange has been referred to as the 'Methodenstreit' or war of methods, a war that at the time seemed to have no clear winners and certainly led to no resolution of the opposing views. Ultimately, of course, Menger's position was far closer to the methodological turn economics took in the subsequent century, although in Germany, Menger's approach and the school that formed around it remained excluded from the university curriculum well into the 20th century.

The vehemence and hostility with which the Germans greeted Menger's *Investigations* is to some degree surprising. Far from an attempt to displace the approach of the Historical School, Menger's *Investigations* is a conscious plan for incorporating many of the features of the historical–empirical approach into a more comprehensive general methodology (Although it must be admitted that Menger's tone is not always cordial when discussing the mistaken views of the Historical School). Menger divides economics into three parts: the historical–statistical which investigates the individual nature and individual connection of economic phenomena, the theoretical which investigates the general nature and general connections of phenomena, and the 'practical sciences of national economy', the basic principles for suitable action in the field of national economy, or in modern terminology, economic policy (*Investigations*, pp. 38–9). Menger defends the idea that science requires knowledge

both of individual (or concrete) aspects of phenomena and of the general (formal) aspects. Presumably, the methods of the Historical School are appropriate to the investigation of concrete aspects of economic phenomena while economic theory is necessary to understand the general aspects. The general form of things, Menger calls *types* and the general form of relationships, Menger calls *typical.*

Menger defends the scientific quality of economic theory despite the fact that its laws are not as strict as some other sciences may be. All sciences, Menger argues, show varying degrees of strictness, and 'the number of natural sciences which absolutely comprise strict laws of nature is also small, and the value of those which only show empirical laws is nevertheless beyond question' (*Investigations*, p. 52). Economic science develops exact laws, but the observation of these laws in reality is hindered by the complexity of the events in which they are manifested and by the impingement of non-economic goals on the actions of observable human beings. Hence, one can never refute the exact laws of economics by pointing to contrary empirical cases. Such a procedure would be analogous to testing the laws of geometry by measuring triangular shapes. In any case, the fact that economic laws are not as strict as some other sciences is irrelevant to its scientific character.

The problem of economic science is to find the causal laws of typical events even though they are manifested in complex reality. Hence it is necessary to 'ascertain the simplest elements of everything real, elements which must be thought of as strictly typical just because they are the simplest' (*Investigations*, p. 60). The appropriate procedure, then, is to start with the simplest elements of economic phenomena and from there investigate the laws by which more complicated human phenomena are formed from simplest elements. Menger called this the 'causal–genetic' approach. Obviously, the simplest elements of economic theory are human valuations and from this can be derived the more complicated economic relationships that are observable in the real world. While Menger does not call this approach 'methodological individualism', it is clear from his discussion of the exact approach and his later criticisms of the excesses of the organic approach that he is a methodological individualist where that means explaining economic phenomena in terms of the choices and consequences of individual human valuation.

Menger's example that he uses to contrast the exact approach with the 'realistic–empirical' is particularly interesting since it clarifies a point of debate about his use of equilibrium constructs. He claims that the exact method can be used to predict 'economic prices' even though one rarely observes true economic prices in the real world. The four criteria for prices to be 'economic' are that (*a*) individuals protect their economic interests completely; (*b*) people have complete knowledge about their goals and their means to achieving them; (*c*) they know the full economic situation (complete knowledge about quantities offered for sale and what prices are being charged) and (*d*) they have the freedom to act in their own interests according to their knowledge. It does not take much imagination to see in these requirements a form of perfect competition where complete knowledge and freedom of entry and exit allow economic man full scope to arrive at equilibrium prices. However, while the laws which predict economic prices are true and exact, the empirical manifestation of them will vary due to circumstances. Indeed, Menger argues that it would be surprising indeed if any of the circumstances required for the establishment of 'economic' prices were ever met completely in the real world. Real prices will deviate from economic prices, and the role of the realistic–empirical approach, then, is to discover to what degree real prices deviate from economic prices. The realistic–empirical approach, however, must take the exact theory of economic prices as the point of departure.

While Menger insists on the necessity of an exact theory of economics to understand economic phenomena, it is clear that he does not believe economics is an all-purpose science. Economics provides exact laws, but only of a subset of human action. In answer to the charge that his vision of human experience is too limited, he emphasizes that a full understanding of social

phenomena requires the aid of the totality of exact sciences of man as well as the historical context of the actions. He is also careful to point out that his assumption of economic man – man guided exclusively by self-interest – is a fiction that does not capture real action. The theory of political economy 'teaches us to follow and understand in an exact way the manifestations of human self-interest in the efforts of economic humans aimed at the provision of their material needs' (*Investigations*, p. 87) but this provides understanding of a special side, by no means the only side, of human life.

One of the criticisms of economic theory that Menger attempted to answer was the charge that pure theory ignored the reality of development and change in economic life and failed to take account of the organic nature of real economic phenomena. While Menger in principle acknowledged the importance of change brought about in time both to empirical forms and to strict types, he believed that the way to explain such change was always with reference to exact theory. In fact, those who discussed organic development missed one of the most important sources of institutional change in social organization. In the *Principles*, Menger had developed a theory of the origin of money as an unintended social order. In the Investigations, Menger generalized his theory to encompass many different social forms. The Historical School's emphasis on historical development required a theory of development, a theory that explained how institutions arise from the unintended consequences of human attempts to improve their own well-being.

Menger saw the problem of exact research to be to discover 'how institutions which serve the common welfare and are extremely significant for its development come into being without a *common will* directed toward establishing them?' (*Investigations*, p. 146). His answer, developed using examples of such social institutions as money, law, language, markets, the origin of communities and of the state itself, was that individuals following their own economic interests provide spillovers to others in the form of increased knowledge of potential advantages or increased ability to pursue their interests. Money, as we have already learned, arises as individuals attempt to overcome the difficulties of barter by acquiring more saleable commodities for the purposes of trade. New localities develop as individuals of different abilities and different professions settle in new areas because they believe they have a better market for their skills. States mostly came into being as families living in close proximity to each other decided it was to their advantage to unite. Most such social organization, Menger argued were not the consequences of conscious planning, but the unconscious result of human will directed toward other, more personal ends. This is the nature of organic development in social science.

What makes Menger's discussion of 'organic' orders (or 'spontaneous orders' as Hayek was later to call them) particularly interesting, is the fact that he not only describes them, but he also provides a brief theoretical analysis of how they can develop. He mentions in his theory of the origin of money that some individuals will be quicker than others to recognize the advantages of acquiring more marketable commodities because it helps them to come closer to their own ends. Not everyone will discover the advantages of indirect exchange at once, but they will soon learn because 'there is no better means to enlighten people about their economic interests than their perceiving the economic successes of those who put the right means to work for attaining them' (*Investigations*, p. 155). It does not take much of a leap to interpret Menger's theory as describing the development of an organic order as a process of discovery and transmission of new information through imitation, motivated by the interests of economic persons. Menger's theory of unintended organic institutions is thus an attempt to reconcile the organic and developmental approach to economics with the exact laws of economic science.

Compared to the frenetic publishing activity of a 20th-century economist, Menger published relatively little during his long career. Nevertheless, he had a major influence on the history of economic thought primarily because he attracted a number of bright and ambitious students. Although his two major disciples, Friedrich Wieser and Eugen Böhm-Bawerk, were never technically his students (both had studied at the

University of Vienna before Menger began teaching there), they were clearly his students in the most important sense: they absorbed and finally extended major aspects of the work of the master. Wieser worked specifically on the problem of imputation which led him to be the first to use the term 'opportunity cost', the utility of the forgone alternative. Wieser also extended Menger's notion of national economy in ways that brought him closer to the to the general equilibrium school. Böhm-Bawerk is best known for his development of Menger's suggestions about the importance of time in production and the implication of goods of higher order for a theory of the structure of production.

While Wieser and Böhm-Bawerk were the best known of Menger's students, there were many others who gathered around him and formed a school. Those who published works in the Austrian tradition included Emil Sax, Johann von Komorzynski, Robert Zuckerkandl, and H. von Schullern-Schrattenhofen. Although not directly his student, Ludwig von Mises (who actually studied under BöhmBawerk) made his first major contribution to economics by extending Menger's notion of marginal utility combined with Menger's process analysis to develop a theory of the value of money. Friedrich Hayek, a student of von Mises, later developed Menger's ideas of spontaneous orders and the problem of knowledge into a comprehensive social theory. Both Mises and Hayek, in turn, have inspired a number of contemporary economists to work in the tradition of Menger to reformulate modern economics in a more 'Austrian' form.

See Also

► Austrian Economics: Recent Work

Selected Works

1871. *Grundsätze der Volkswirtschaftslehre*. Trans. J. Dingwall and B.F. Hoselitz as *Principles of economics* by with an Introduction F.A. Hayek. New York/London: New York University Press, 1981.

1883. *Untersuchungen über die Methode der Sozialwissenschaften und der politischen Ökonomie insbesondere*. Trans. F.J. Nock as *Problems of economics and sociology*, edited and with an Introduction by L. Schneider Urbana: University of Illinois Press, 1963. Reprinted as *Investigations into the method of the social sciences with special reference to economics* with a new Introduction by L.H. White, New York/London: New York University Press, 1985.

1884. *Irrthümer des Historismus in der deutschen Nationalökonomie.* Vienna: Hölder.

1887. *Zur Kritik der politischen Ökonomie*. Vienna.

1888. Zur Theorie des Kapitals. *Conrad's Jahrbücher für Nationalökonomie und Statistik* 17: 1–49.

1889. Grundzüge einer Klassification der Wirtschaftswissenschaften. *Conrad's Jahrbücher für Nationalökonomie und Statistik* 14.

1892. Geld. In *Handworterbuch der Staatswissenschaften*, vol. 3. Vienna.

1892a. Die Valutaregulierung in Österreich-Ungarn. *Conrad's Jahrbücher fur Nationalökonomie und Statistik* 3.

1892b. *Der Übergang zur Goldwahrung. In Untersuchungen über die Wertprobleme der österreichisch-ungarischen Valutareform*. Vienna.

1892c. La monnaie mesure de la valeur. *Revue d'économie politique* 6.

1892d. On the origin of money. *Economic Journal* 2(6): 239–255.

Bibliography

Alter, M. 1982. Carl Menger and homo oeconomicus: Some thoughts on Austrian theory and methodology. *Journal of Economic Issues* 16 (1): 149–160.

Bloch, H.-S. 1940. Carl Menger: The founder of the Austrian school. *Journal of Political Economy* 48: 428–433.

Hayek, F.A. 1981. Carl Menger. Introduction to Carl Menger, *Principles of economics*. New York/London: New York University Press.

Hicks, J.R., and W. Weber, eds. 1973. *Carl Menger and the Austrian school of economics*. Oxford: Clarendon Press.

Jaffé, W. 1976. Menger, Jevons and Walras dehomogenized. *Economic Inquiry* 14: 511–524.
Kauder, E. 1957. Intellectual and political roots of the older Austrian school. *Zeitschrift für Nationalökonomie* 17: 411–425.
Kauder, E. 1959. Menger and his library. *Economic Review* 10 (Hitotsubashi University).
Kauder, E. 1965. *A history of marginal utility theory.* Princeton: Princeton University Press.
Kirzner, I.M. 1979. The entrepreneurial role in Menger's system. In *Perception, opportunity and profit: Studies in the theory of entrepreneurship*, ed. I.M. Kirzner. Chicago: University of Chicago Press.
Martin, D.T. 1979. Alternative views of Mengerian entrepreneurship. *History of Political Economy* 11: 271–285.
Schumpeter, J.A. 1951. Carl Menger, 1840–1921. In *Ten great economists, from Marx to Keynes*, ed. J.A. Schumpeter. New York: Oxford University Press.
Stigler, G.J. 1941. Carl Menger. In *Production and distribution theories*, ed. G.J. Stigler. New York: Macmillan.
Streissler, E. 1972. To what extent was the Austrian school marginalist? *History of Political Economy* 4: 426–441.
von Mises, L. 1978. Carl Menger and the Austrian school of economics. In *The clash of group interests and other essays*, ed. L. von Mises. New York: Center for Libertarian Studies.
Wagner, R.E. et al. 1978. Carl Menger and Austrian economics. *Atlantic Economic Journal* 6(3), Special issue. Contributions by R.E. Wagner, S. Bostaph, L.S. Moss, I. Kirzner and H. Nelson Gram and V.C. Walsh, L.M. Lachmann and K.I. Vaughn.

Bibliographic Addendum

See also Alter, M. 1990. *Carl Menger and the origins of Austrian economics.* Boulder: Westview Press, Different facets of Menger's work are discussed in B. Caldwell, ed., *Carl Menger and his legacy in economics.* Durham: Duke University Press, 1991. M. Latzer and S. Schmitz, eds., *Carl Menger and the evolution of payments systems.* Cheltenham: Edward Elgar, 2002, provides the first English translation of Menger's 'Geld' as well as essays both evaluating Menger's views on monetary systems and applying them to contemporaneous issues.

Menger, Karl (1902–1985)

G. Schwödiauer

Karl Menger was born in Vienna on 13 January 1902, the son of Carl Menger, the founder of the Austrian school of economics, and died in Chicago on 5 October 1985. He studied mathematics, physics and philosophy at the University of Vienna from 1920 to 1924, where he received his doctoral degree in mathematics. In 1925 he went to Amsterdam (as an assistant of L.E.J. Brouwer) where he continued his research on the theory of curves and dimension theory which led to his habilitation as docent in 1926. In 1927, Menger was appointed associate professor of geometry at the University of Vienna, in which position he remained, interrupted by visiting professorships at Harvard University and the Rice Institute (1930/31), and at the University of Notre Dame (1937/38), until 1938.

Due to his strong philosophical interests he joined the so-called Vienna Circle of logical-empiricist philosophers founded by the philosopher M. Schlick and Menger's former teacher, the mathematician H. Hahn. His most remarkable students at that time were K. Gödel and A. Wald. Menger organized a mathematical colloquium of his own the proceedings of which were published as *Ergebnisse eines Mathematischen Kolloquiums'* (Vienna, 1931–7). Menger's colloquium provided a forum not just for original contributions to logic and pure mathematics but also for rigorous investigations into fundamental problems of various empirical sciences, among them economics: it was in Menger's colloquium where Wald presented for the first time his path-breaking proof of the existence of a Walrasian competitive equilibrium, and where von Neumann read his paper on the equilibrium of an expanding economy (first published in vol. 8 of the *Ergebnisse* . . ., 1937).

When Hitler occupied Austria in March 1938, Menger, who at that time was teaching at Notre Dame, immediately resigned from his professorship in Vienna. He accepted a permanent position at the University of Notre Dame where he edited the *Reports of a Mathematical Colloquium, 2nd series* (1937–46). From 1946 to his retirement in 1971 he was professor of mathematics at the Illinois Institute of Technology, Chicago. During these years and after he was offered and accepted visiting professorships at various European university institutes (among them the Sorbonne in Paris, and the Institute for Advanced Studies in Vienna).

Karl Menger was a creative mathematician who made important contributions to pure and applied mathematics as well as to its logical and philosophical foundations. This is, however, not the place to evaluate his achievements in his main field of research. Here we have to confine ourselves to the impact Menger's extraordinarily broad and penetrating intellect had on the social sciences in general, and on economic theory in particular. Menger was, with the exception of Schlick, the only member of the Vienna Circle seriously interested in the study of ethical problems. In this book *Moral, Wille und Weltgestaltung*, published in 1934 (English edition 1974, under the title *Morality, Decision and Social Organization*) Menger applied rigorous logico-deductive thinking to the field of ethics strictly avoiding any value judgements and metaphysical arguments. He arrived at the conclusion that the only basis of a person's conduct is the person's decisions. In particular, Menger rejected Kant's categorical imperative for failing to provide a basis for the regulation of a person's conduct, by demonstrating that generically several mutually incompatible types of behaviour are compatible with Kant's principle. Menger's positive approach to ethics is based on the 'externalization of ethics', i.e. the association of the groups of its adherents to a norm, a moral code, or any value judgement, and the study of the relations between the groups of its adherents. It turns out that Kant's categorical imperative is not only not a sufficient but also not a necessary condition for compatibility and the constitution of cohesive groups. Menger also points to the potentially fruitful role of controlled social experimentation for scientific ethics. Though more or less neglected by his contemporaries and forgotten later on, Menger's work on ethics and social organization (including his papers 'Einige neuere Fortschritte in der exakten Behandlung sozialwissenschaftlicher Probleme' (1936) and 'An Exact Theory of Social Groups and Relations' (1938)) has to be considered a pioneering contribution to the rigorous modelling of social decision-making problems.

From his youth, when he edited and introduced the second German edition (1923) of his father's Principles of Economics, to his late life (cf. his 1972/73 paper on Austrian marginalism) Menger had been deeply interested in economic theory.

His own contribution is concerned with the role of uncertainty in economics and with the law of diminishing returns. Menger's essay on uncertainty, which was published in German in 1934 (but was according to his own account, essentially completed by 1923 and presented to the Vienna Economic Society in 1927), (deals with the well-known problem called Bernoulli's paradox, but instead of trying to solve it by the introduction of a concave utility function for wealth, Menger focused on subjective probability and its empirical estimation. Though different in outlook, it stimulated (as Morgenstern reports) von Neumann's axiomatic treatment of utility theory. Menger's 1936 papers on the law of returns were called by himself a study in meta-economics. At that time a variety of formulations and alleged proofs of this law were available from economic literature. Menger not only cleared up the logico-mathematical status and interrelationships of the versions given by Böhm-Bawerk, Wicksell and Mises, but provided a firm basis for the further mathematical study of the properties of production functions.

Selected Works

1934a. *Moral, Wille und Weltgestaltung*. Vienna: Springer. Trans. with a postscript by the author as *Morality, decision and social organization*. Dordrecht/Boston: Reidel, 1974.

1934b. Das Unsicherheitsmoment in der Wertlehre. Betrachtungen in Anschluss an das sogenannte Petersburger Spiel. *Zeitschrift für Nationalökonomie* 5: 459–485.

1934c. Bemerkungen zu 'Das Unsicherheitsmoment in der Wertlehre'. *Zeitschrift für Nationalökonomie* 6: 283–285. Of 1934b and 1934c Trans. as: 'The role of uncertainty in economics'. In *Essays in mathematical economics in honor of Oskar Morgenstern*, ed. M. Shubik. Princeton: Princeton University Press, 1967.

1936a. Bemerkungen zu den Ertragsgesetzen. *Zeitschrift für Nationalökonomie* 7: 25–26.

1936b. Weitere Bemerkungen zu den Ertragsgesetzen. *Zeitschrift für Nationalökonomie* 7: 388–397. 1936a and 1936b Trans. as 'The logic of the laws of return. A study in meta-economics'. In *Economic activity analysis*, ed. O. Morgenstern. New York: Wiley, 1954.

1936c. Einige neuere Fortschritte in der exakten Behandlung sozialwissenschaftlicher Probleme. In *Neuere Fortschritte in den exakten Wissenschaften. Fünf Wiener Vorträge. Dritter Zyklus*. Leipzig: Deuticke, 1936.

1938. An exact theory of social groups and relations. *American Journal of Sociology* 43: 790–798.

1972. Österreichischer Marginalismus und mathematische Ökonomie. *Zeitschrift für Nationalökonomie* 32: 19–28. Trans., expanded by the author, as 'Austrian marginalism and mathematical economics'. In *Carl Menger and the Austrian School of Economics*, ed. J.R. Hicks and W. Weber. Oxford: Oxford University Press, 1973.

1979. *Selected papers in logic and foundations, didactics, economics*. Dordrecht: Reidel. (Contains translations of 1934b, 1934c, 1936a, 1936b.)

Mercantilism

Laura LaHaye

Abstract

Mercantilism is economic nationalism that seeks to limit the competition faced by domestic producers. It refers to the economic thought and policies that were characteristic of the dominant Western European trading nations during the transition from feudalism to modern capitalism from the 16th to the late 18th century. It is often depicted as the school of thought that confused money with wealth, promoting a favourable balance of trade as the best method to increase the wealth of a nation that did not possess gold or silver mines.

Keywords

Colbert, Jean-Baptiste; East India Company; German Historical School; Heckscher, E. F.; Mercantilism; Money; Monopoly; Mun, T.; Navigation acts; Protection; Specie; Tariffs

JEL Classifications

B11

Mercantilism is economic nationalism that seeks to limit the competition faced by domestic producers. The tools of mercantilist policies include the granting of monopoly privileges, regulation of prices and business practices and especially prohibitions, tariffs, subsidies and other regulations regarding the conduct of international trade. The goals of mercantilism are supposedly to contribute to the development of a rich and powerful state; however, the principal beneficiaries are the merchants and producers who are protected or encouraged under a mercantile system. Although mercantilism was frequently promoted as means of obtaining longterm development objectives, it is significant that such promotion typically increased in fervour following periods of trade crisis, such as that in England in the 1620s.

Mercantilism refers to the economic thought and policies that were characteristic of the dominant western European trading nations during the transition from medieval feudalism to modern capitalism from the 16th to the late 18th century. Adam Smith (1776, p. 399) characterized the 'principle of the commercial or mercantile system' – that a 'favourable' balance of trade would bring gold or silver into the country – which could be used to 'carry on foreign wars, and to maintain fleets and armies in distant countries'. Import restraints and encouragement to exportation were the mercantile policies that would enrich and empower the newly emerging nation-states. At the end of the 19th century authors of the German Historical School popularized the term 'mercantilism' while rationalizing the mercantile policies as necessary for the

unification of feudal power centres by large competitive states.

The mercantile era emerged following the discovery of the New World and the East Indies by European explorers at the close of the 15th century. Shipping and trading grew in importance during this period as did the frequency of military battles at sea and in the colonies. Anglo-French rivalry remained intense, and Henry VIII invested heavily in shipping while fortifying the coastline against possible attack. Meanwhile, the Spanish Habsburgs were at war all over Europe. Mercantile economic warfare complemented the military objectives of the antagonistic nations and served to unify each nation against an external threat.

As a concept of society, mercantilism reflects the medieval view that wise government intervention is necessary to delicately balance the tendencies of unbridled competition to produce unjust wages or income below a subsistence level, when too many workers or businesses operate in a particular activity, or to result in an unregulated monopoly that would reap unjust profits charging prices that are too high. The market could certainly not be left to itself to find a 'just price' or wage.

The 1563 Statute of Artificers marked one of the first efforts by Queen Elizabeth of England to extend the restrictive and regulatory policies of medieval towns to the nation as a whole. A century later, Louis XIV of France, with the assistance of his powerful mercantilist finance minister Jean-Baptiste Colbert, undertook similar national regulation of industry and simplification of the internal tolls of France which Heckscher (1935, vol. 1, p. 103) 'ranks with Elizabeth's Statute of Artificers as one of the two unquestionable triumphs of mercantilism in the sphere of economic unification'.

The granting of monopoly privileges was a relatively more important form of state protection during the earlier part of the mercantile era. The British East India Company was granted a monopoly charter by Queen Elizabeth in 1600 which encouraged the United Provinces to consolidate the independent Dutch traders into the Dutch East India Company in 1604. A number of short-lived East Indies trading monopolies were chartered by the French Crown throughout the first half of the 17th century, culminating in the 1664 charter of, and royal participation in, the French East Indies Company. These privileges were intended to benefit the developing shipping and long-distance trading industries themselves as well as to provide revenues to the state either directly, in the case of state monopolies, or indirectly through modest duties on imports of the private monopolies.

When, however, the successful conclusion of the Dutch Revolt in 1648 exposed the English to an increased level of competition in intra-European shipping and trading, Cromwell eventually responded with the first Navigation Act of 1651. This Act stipulated that all goods imported into England or her territories had to be carried in English ships, unless they were carried directly from a European country of origin on ships owned and crewed by citizens of that country of origin, and that no foreign vessels could engage in the coastal trade among English ports. Furthermore, no type of salted fish or fishing by-product of the type usually caught and processed by English people could be imported unless it was caught and processed by an English ship. Additional navigation laws further protected English fishing, shipping and trading industries from competition, especially from the Dutch, who largely dominated maritime activity at the time.

More general industrial protection followed the navigation laws, although several early examples of discriminatory protective policies were already in existence. The 1667 anti-Dutch tariff imposed by Colbert in France, and the subsequent quadrupling of import duties in England during the 15 years following the 1688 accession to the throne of William III and Mary marked the major shift from moderate revenue-generating customs duties on imports and exports to the more protective import tariffs as well as bounties and drawbacks on exports that constituted the mercantile system in Smith's view. English export duties on woollens were abolished in 1700 and export duties were abolished in general by the Walpole customs reform of 1722. Protection was further extended throughout the 18th century until 'the building up of the protective system showed signs of becoming a general and recognized policy ... in

the decade in which Adam Smith was collecting material and writing his great blast against commercial regulation, *The Wealth of Nations'* (Davis 1966, p. 314).

Following Smith's (1776, p. 418) lengthy examination of the 'popular notion that wealth consists in money', mercantilism has often been depicted as the school of thought that confused money with wealth. Although this interpretation has been thoroughly debated, there is certainly much evidence to suggest that mercantile pamphleteers did believe an inflow of precious metals would increase the wealth of the nation and that foreign but not domestic or internal trade was the only way to increase the wealth of a nation that did not possess gold or silver mines. Exportation of bullion or coin had generally been regulated or prohibited since medieval times, and it was in an effort to get those restrictions relaxed that mercantilist authors such as Mun (1664, p. 5), a director of the British East Indies Trading Company, argued that the 'means therefore to increase our wealth and treasure is by *Foreign Trade,* wherein we must ever observe this rule; to sell more to strangers yearly than we consume of theirs in value'. That the wealth of the nation was not perceived to be primarily related to its ability to provide goods and services to its consumers is revealed when reading Mun's (1664, p. 7) recommendations for reducing imports such as using waste grounds 'to supply our selves and prevent the importations of Hemp, Flax, Cordage, Tobacco and divers other things which we now fetch from strangers to our great impoverishing'.

In all fairness, the proponents of an export surplus did not generally advocate the accumulation of specie for the simple purpose of hoarding it, although they did like to make the analogy between the kingdom and an individual that would grow poor if its purchases exceeded its income. Of course, neither the individual nor the kingdom will grow poor if the purchases include investment expenditures that yield a rate of return in excess of the borrowing cost. As a store of value, money is only a component of wealth to the extent that one intends to spend it 1 day, and there is a limit to this precautionary motive for accumulating specie. It is sensible to accumulate specie following a period of declining reserves (excessive expenditure) or in response to increased uncertainty, which requires a larger precautionary balance, or in response to increased hostility, which requires a larger defence balance, but not ad infinitum, except perhaps to maintain a desired ratio of specie to growing royal expenditures over time. For the merchant adventurers engaged in long-distance trading, specie was a valuable factor of production as a medium of exchange, and they recognized the relationship between the quantity of money in circulation and the amount of trading activity that could be financed. Mun (1664, p. 68) was careful to recommend that the royal treasure should not be augmented by more than the favourable balance of trade, 'for if he should mass up more money than is gained by the over-balance of his foreign trade, he shall not *Fleece,* but *Flea* his Subjects, ... whereby the life of lands and arts must fail and fall to the ruin both of the public and private wealth'. This indicates that he perceived a relationship between the quantity of money and the level of national economic activity, although his immediate concern was probably the economic activity of his own British East India Company, which imported exotic goods that could not be produced at home.

More important, perhaps, than enabling the royal treasure to be augmented, an export surplus is generally perceived to stimulate domestic employment directly or indirectly by reducing interest rates. According to Heckscher (1935, vol. 2, p. 121), the '"fear of goods" was nourished ... by the idea of creating work at home and of taking measures against unemployment'. References to the unemployment argument date back to the early 15th century, and in English legislation in 1455, 'foreign competition was blamed for having caused the unemployment in the silk industry' (Heckscher 1935, vol. 2, p. 122). The preference for encouraging exportation of manufactured consumer goods, as opposed to raw materials or productive equipment, and allowing the importation of raw materials are consistent with this employment concern. An export surplus – an excess of domestic saving over investment – naturally arises when productivity growth outpaces the growth of profitable domestic investment opportunities, and this may include an accumulation of international reserves

to finance the growth of monetized transactions; but to try to engineer such a surplus with protective trade policies would be futile at best. In addition to competitively induced innovation and increased specialization, limited by the extent of the market, productive investment is the true source of a sustainable increase in the wealth of a nation, and there is no reason to suppose, a priori, that domestic investment is inferior to foreign investment.

Most of the vestiges of the mercantile era were removed during the laissez-faire era of the 19th and early 20th centuries, especially in England, where monarchical power was weaker and property rights were clearer than in France and Spain. Yet mercantilism has remained a topic of considerable debate, especially since Heckscher's broad treatment of the subject and the emergence of global depression in the 1930s (Heckscher 1935; Viner 1937; Minchinton 1969; Coleman 1969; Magnusson 1993). Whether mercantilist policies re-emerge in the 21st century will depend on the institutional framework within which the special interests seeking protection must function (Ekelund and Tollison 1997), as there exists no coherent economic doctrine to support such policies.

See Also

- ▶ Cameralism
- ▶ Colbert, Jean-Baptiste (1619–1683)
- ▶ Hume, David (1711–1776)
- ▶ Misselden, Edward (fl. 1608–1654)
- ▶ Mun, Thomas (1571–1641)
- ▶ Schmoller, Gustav von (1838–1917)

Bibliography

Coleman, D. 1969. *Revisions in mercantilism*. London: Methuen.

Davis, R. 1966. The rise of protection in England, 1689–1786. *The Economic History Review* 19: 306–317.

Ekelund, R., and R. Tollison. 1997. *Politicized economies: Monarch, monopoly and mercantilism*. College Station: Texas A & M University Press.

Heckscher, E. 1935. *Mercantilism* . London: Allen & Unwin.2 vols

LaHaye, L. 2007. Mercantilism. In *The concise encyclopedia of economics*, ed. D. Henderson. Indianapolis: Liberty Fund.

Magnusson, L. 1993. *Mercantilist economics*. Boston: Kluwer.

Magnusson, L. 1994. *Mercantilism, the shaping of an economic language*. London: Routledge.

Minchinton, W. 1969. *Mercantilism, system or expediency?* Lexington: Raytheon.

Mun, T. 1664. England's treasure by foreign trade or the balance of our foreign trade is the rule of our treasure. London. Reprints of Economic Classics, New York: Kelley, 1968.

Smith, A. 1776. In *The wealth of nations*, ed. Edwin Cannan. New York: Random House. 1937.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.

Mercier De La Rivière, Pierre-Paul (Mercier or Lemercier) (1720–1793/4)

Peter Groenewegen

Keywords

Balance of trade; Mercier de la Rivière, P. P.; Money; Physiocracy; Private property; Taxation theory; Value theory, Physiocratic

JEL Classifications

B31

Lawyer, administrator and economist, born into a financier's family in 1720. From 1749 to 1759, he was Councillor of the Paris Parlement; from 1759 to 1764, Governor of Martinique. Although Garnier (1854, p. 188) claims that Mercier became acquainted with Quesnay and Mirabeau while Governor of Martinique, this is doubtful. However, after 1765 he became a prominent Physiocrat and published what many (for example, Smith 1776, p. 679; Mill 1824, p. 712) considered to be the most comprehensive exposition of Physiocratic doctrine in his *L'ordre naturel et essentiel des sociétés politiques* (1767). This gained him both Catherine the Great's invitation to advise her on a new legal code and the

enmity of Voltaire (1768), who devastatingly satirized his cumbrous prose. Du Pont (1768) wrote a summary of Mercier's work for *Ephémérides,* confirming thereby its enormous importance for the Physiocrats. Subsequently, Mercier published a reply to Galiani's dialogues attacking the Physiocratic position on the grain trade (1770) and an essay on the importance of public education dedicated to the King of Sweden (1775). He died in Paris in either 1793 or 1794.

Mercier's *L'ordre naturel* (1767) is therefore the major general treatise of Physiocratic doctrine both political and economic. The work divides into three parts with a concluding summary chapter. Part I develops the theory and necessity of the social order based on the duties and rights inherent in private property, without which a society cannot be sustained. 'The greatest possible happiness comes from the greatest possible abundance of means of enjoyment and the greatest possible freedom to profit from [the ownership of property]' (1767, I, pp. 42–3). Hence the sanctity of private property and complete freedom for its owners to use it are the first principles of the theory of natural order (pp. 45, 50–51). These principles need to be inculcated in society through a system of public instruction (pp. 91–2). Part II discusses the manner in which social order is achieved in practice through the establishment of three fundamental institutions: law and magistrature, the sovereign as bearer of authority, and institutions of public instruction for spreading knowledge of the social order among all members of society. In his lengthy elaboration on these institutions (chs. 11–24) Mercier presents his famous defence of legal despotism.

Part III (the greater part of volume 2 in the original edition) further discusses the practical promotion of the social order by examining the political economy of wealth creation. After reviewing the essential association between the king and his subjects (ch. 25) the theory of taxation is presented as the way in which kings share the net product of their common property with the landlords (chs. 28–34). The dogmatic presentation of Physiocratic tax theory was the special target of Voltaire (1768). These chapters also contain interesting economic contributions. In them Mercier emphasizes the role of consumption and effective demand in stimulating reproduction (vol. 2, pp. 138–9); presents an argument showing the possibility of a downward spiral in economic activity 'in geometrical progression' if taxation reduces the advances of agriculture (pp. 150–1), an analysis having both real and value aspects (pp. 160–4). The second half of Part III examines commerce and industry and their function in the Physiocratic social order (chs. 35–43), the last chapter being a particularly dogmatic demonstration of these activities' unproductive nature. However, they likewise contain interesting analytical contributions on the role of money and its circulation (vol. 2, pp. 262–3, 297–9, 334), the impact of trade on wealth via the profits of agriculture and hence accumulation when it provides a wider market for agricultural produce (p. 273) and a critique of the balance of trade doctrine based on the logical impossibility for all nations to enjoy a favourable balance (p. 349) and a type of specie mechanism argument (pp. 360–7) from which Mercier concludes that nations can have too much as well as too little money (pp. 368–9). His discussion of commerce and industry highlights, in particular, the richness of Physiocratic value theory and its importance for their theory of distribution and economic development. As Vaggi (1987) has demonstrated, recognition of this importance is indispensable for a proper understanding of Physiocracy, as is the full social and political framework in which their policy recommendations are framed and for which Mercier was particularly noted by his contemporaries.

Selected Works

1767. *L'ordre naturel et essentiel des sociétés politiques*. London/Paris.

1770. *L'intérêt général de l'état, ou la liberté du commerce des blés, avec une réfutation d'un nouveau système publié par l'abbé Galiani en forme de dialogues sur le commerce des blés*. Amsterdam/Paris.

1775. *De l'instruction publique, ou considérations morales et politiques sur la nécessité, la nature et la source de cette institution*. Stockholm/Paris.

Bibliography

de Voltaire, F.M.A. 1768. *The man of forty crowns.* Trans. from the French. London.
Du Pont de Nemours, P.S. 1768. De l'origine et des progrès d'une science nouvelle. Reprinted in *Physiocrates*, vol. 1, ed. E. Daire. Paris, 1846.
Garnier, J. 1854. Mirabeau. In *Dictionnaire de l'économie politique,* vol. 2, ed. Paris: Ch. Coquelin and Guillaumin.
Mill, J. 1824. Economists. Supplement to *Encyclopaedia Britannica,* vol. 3. Edinburgh.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell, A.S. Skinner, and W.B. Todd. Oxford: Clarendon Press. 1976.
Vaggi, G. 1987. *The economics of François Quesnay.* London: Macmillan.

Mercosur

Marcelo Olarreaga

Abstract

Mercosur is an ambitious economic integration project, launched in 1991, which includes Argentina, Brazil, Paraguay and Uruguay. The early and quasi-complete liberalization of intra-regional trade and the adoption of a common external tariff by 1996 were accompanied by significant increases in intra-regional trade. However, the most difficult and challenging steps towards a common market (its original objective) has been slow since then, in part due to the absence of strong regional institutions.

Keywords

Andean community; Central-bank independence; Common external tariffs; Comparative advantage; Customs unions; Foreign direct investment; Free-trade agreements; Intra-regional trade; Mercosur; Monetary union; Non-tariff barriers; Tariffs; Trade diversion; World Trade Organization

JEL Classifications

F1; F15

Mercado Común del Sur (Mercosur, Southern Common Market) is an ambitious economic integration project which includes Argentina, Brazil, Paraguay and Uruguay. It represents 70 per cent of the gross domestic product (GDP) of South America and 60 per cent of its population. In terms of geographic size, Mercosur is four times larger than the European Union, which would rank Mercosur as the largest customs union in the world. Its economic size, however, is similar to that of the Netherlands.

Mercosur was launched in March 1991 with the signing of the Asunción Treaty. Aiming at creating a common market, Article I calls for full internal mobility of goods, services and factors of production, the implementation of common external policies in these areas, as well as the coordination of macroeconomic policies and cooperation in education, health and transport policies.

It is an agreement that is open to accession by all members of the Latin American Integration Association (which regulates partial bilateral trade agreements among members). By 1996 Bolivia and Chile were associate members of Mercosur; later, a free-trade agreement (FTA) was signed with the Andean Community. At the time of writing other Latin American countries are in different stages of association with Mercosur. Negotiations for trade agreements are ongoing with China, the European Union, Mexico, India, South Africa, Egypt, and Morocco.

Mercosur members had agreed in the Asunción Treaty to create the common market within four years. However, this proved politically impossible and little progress was made a part from very rapid reductions in internal tariffs (with some negotiated exceptions). Very quickly it became clear that the ambitious objectives of the Asunción Treaty had to be scaled back. An imperfect customs union became a more realistic objective, and the Protocol of Ouro Preto signed in December 1994 called for the implementation of a common external tariff (CET) by early 1995. It was an imperfect 'common' external tariff as each member was allowed some deviations from the negotiated CET; and more than ten years later the CET is still to be defined in some politically entrenched sectors (such as sugar). Nevertheless, by 1996 internal

tariffs were applied on less than three per cent of tariff lines, and the CET was implemented in 80 per cent of tariff lines.

In all other areas progress has been slow or non-existent. For example, non-tariff barriers (NTBs) are not only not subject to common external policies but are routinely used as an impediment to intra-regional trade, contrary to what is explicitly required in Article V of the Asunción Treaty. For example, non-automatic import licensing, sanitary measures and other technical regulations (such as labelling) on Brazilian imports of powdered milk impose an equivalent tax of 54 per cent on Argentina's exporters (Berlinski 2004). Internal trade in the automobile sector is managed with bilateral trade quotas at the firm level (for those firms with a presence in several Mercosur members) and a trade balance constraint on global automobile trade, which if removed could double bilateral trade (Brambilla 2005). Negotiations on services trade and factor mobility were still at a very early stage 15 years after the treaty was signed. The Services Trade Protocol signed in 1997 merely states the multilateral commitments of Mercosur members at the World Trade Organization (WTO). The dispute settlement mechanism (DSM) remained unused until 1997; an appeal court was created only in 2002. Steps have been taken for the mutual recognition of standards, but enforcement has been largely absent (for example, in the area of education, mutual recognition stops at the high-school level). Macroeconomic coordination is limited to routine exchange of (public) information.

Internal Tariffs and the FTA

In spite of the slow progress in the 'non-tariff' areas (NTBs, services, factor mobility, macroeconomic coordination), by the late 1990s Mercosur was considered one of the most successful attempts at regional integration between developing countries. This was partly due to the unprecedented rapid elimination of internal tariffs, a sixfold increase in intra-regional trade, a twentyfold increase in flows of foreign direct investment (FDI) (mainly from the United States and Spain), and the longevity that the agreement was achieving in spite of several financial, economic and institutional crises.

A more careful analysis, however, reveals a more subtle picture. Let me start with the rapid increase in trade. Yeats (1998) argued that intra-regional trade appeared to be concentrated in products in which Mercosur did not have a clear comparative advantage (capital goods), and that these were the goods with the most rapid growth after the creation of Mercosur. He concluded that this provided evidence of trade diversion and should raise questions regarding the (static) welfare impacts of such rapid growth in intra-regional trade. Olarreaga and Soloaga (1998) showed that fast-growing intra-regional trade was concentrated on products with trade-diverting potential partly because deviations from zero internal tariffs occurred in products with substantial trade-creation potential, as predicted by the theoretical political economy literature on regional agreements.

External Tariffs and the CET

It has also been argued that a significant part of the increase in intra-regional trade need not be attributed to the creation of Mercosur, but rather to the tremendous trade liberalization vis-à-vis the rest of the world that Mercosur members were independently undertaking after the mid-1980s. For example, Brazil's external tariff declined from an average of 80 per cent in the mid-1980s to an average of 15 per cent by the mid-1990s. This can explain a large share of the rapid growth in imports, including those from other Mercosur members. On the other hand, it has been suggested that the important external liberalization undertaken by Mercosur members needs to be partly attributed to the creation of Mercosur. Without the significant competitive pressure imposed by the increase in intra-regional flows, the move towards lower external tariffs would have been more difficult. Bohara et al. (2004) showed that the lobbying for high external tariffs was eroded by the increase in intra-regional trade due to internal tariff preferences. Also, it has been shown that a significant force for lower CET was the prospect of the elimination of duty drawbacks for intra-

regional exports (a by-product of the creation of Mercosur) as agreed in Ouro Preto. Indeed, the elimination of duty drawbacks on intra-regional exports increased counter-lobbying by regional exporters for lower tariffs on their imports of intermediate inputs from the rest of the world. This led to a 25 per cent reduction in the negotiated CET (Cadot et al. 2003).

An additional trade-related benefit for Mercosur members is that rest-of-the-world exporters to the regional market started pricing their products more competitively due to the more intense competition in the internal market brought by tariff preferences granted to other Mercosur members. This led to significant welfare gains for Mercosur consumers of imported products at the expense of foreign firms exporting to the region (Chang and Winters 2002). Schiff and Chang (2003) further showed that the pro-competitive forces that led rest-of-the-world exporters to price more competitively after the creation of Mercosur were also present even when Mercosur partners did not export to each other, as long as they had the potential to do so (that is, markets were contestable). Thus, Mercosur created trade-related gains to its members even in the absence of any intra-regional trade flow or external tariff reduction.

Beyond Tariffs

Moreover, regardless of whether Mercosur led to trade diversion, it has been shown that most households and in particular poor households within the region benefited from the agreement. Porto (2006) provided evidence of a pro-poor bias of Mercosur in Argentina: on average, poor households gain more from the reform than middle-income households, whereas the effects on rich families are positive but not statistically significant. Prior to Mercosur, Argentine trade policy protected the rich over the poor. As relative pre-Mercosur tariffs are higher on relatively skill-intensive goods, the tariff removals tend to benefit the poor over the rich. Thus, Mercosur not only helps reduce poverty in Argentina, but it improves the distribution of income.

Regarding the rapid increase in FDI flows, it seems that the creation of Mercosur was not the main cause. Most statistical analysis shows no direct causality between the creation of Mercosur and the rapid growth in FDI (Castilho and Zignago 2002). The main forces were the simultaneous privatization processes in Argentina and Brazil, the macroeconomic stabilization and the external tariff reduction independently undertaken by Mercosur members, which provided foreign firms investing in the region access to imported inputs (Chudnovski 2001). The creation of a larger regional market only marginally contributed to foreign firms' decisions to invest in Mercosur.

The longevity of Mercosur has come at the cost of achievements in the area of internal free trade, the implementation of the CET, and an (implicit) consensus to move slowly in other areas. For example, at the end of 1992 Argentina increased its statistical import tax surcharge (applied to intra-Mercosur imports) from three to ten per cent as its trade deficit with Brazil widened. An optional increase in external tariffs of up to three percentage points was authorized in 1997. In June 2001, on the eve of a major financial and fiscal crisis, the Argentine government unilaterally altered its tariff rates on capital goods and consumer goods. A waiver was granted by the Common Market Council. In 2006, duty drawbacks and temporary admission regimes which were to be eliminated by 2000 were still in place; the customs code drafted in 1994 had not been adopted by any of the members' parliaments; and no common safeguard mechanism had been put in place to deal with unforeseen changes in competitive pressures, leading to the adoption of unilateral ad hoc measures and private sector marketing agreements (for example, dairy, paper and steel) after the devaluation of the real in January 1999. In sum, flexibility rather than consistency has been the norm, and time-inconsistent policies have often been reversed with the associated cost for the credibility of Mercosur institutions (Bouzas 2002).

Regional Institutions

From the very beginning Mercosur decisions were driven by national private-sector interests, and weak and relatively politicized regional institutions emerged, partly because Brazil (the largest

member) wanted to preserve its hegemony. Mercosur is ruled by a Consejo del Mercado Común (CMC, Common Market Council), which is responsible for the political decisions of the integration process. Sitting members are the four national presidents and their cabinets, who regularly meet twice a year. The Grupo Mercado Común (GMC, Common Market Group) is directly answerable to the CMC and is the executive organ, which includes the ministers of foreign affairs and economics, the chairmen of the central banks, and the permanent coordinators from each member country. The GMC enforces resolutions. The GMC branches out into the Trade Commission of Mercosur, which is responsible for counselling and enforcing trade policy instruments as well as setting directives; the Joint Parliamentary Commission in representation of the four parliaments; the Economic and Social Consultation Forum, which has representatives from the different economic and social groups; and finally a weak Administrative and Technical Secretariat, which supports the whole operation from Montevideo. With such a structure, any decision is likely to be highly politicized (Vaillant 2005).

The absence of strong regional institutions has been particularly felt in the area of macroeconomic coordination. Throughout the 1990s the variability of nominal exchange rates within the region was twice as great as in other comparable countries, leading to the strong backlashes against regional integration discussed above. This led some regional leaders, including the former Argentine President Carlos Menem, to call for the creation of a monetary union. As argued by Eichengreen (1998), this is the optimal instrument to avoid wide fluctuations in intra-regional exchange rates while keeping some flexibility with respect to bilateral exchange rates with the rest of the world. However, a monetary union will not be an option as long as the other institutions of Mercosur remained politicized and weak. As the experience of the European Union shows, a monetary union not only requires a strong and politically independent central bank but should also be part of an interlocking web of strong economic and political agreements, all of which could be jeopardized if a country abandoned the single currency. The latter acts as a significant barrier to exit for members, reinforcing credibility and stabilizing markets. Mercosur members have all taken significant steps towards central-bank independence. But in terms of barriers to exit there is not much apart from a relatively well-functioning customs union. Very little has been achieved in terms of common trade, economic, social or security policies. If Mercosur does not engage in a deeper integration project, a monetary union cannot be successful.

To conclude, Mercosur is an unprecedented example of successful and enduring regional integration among developing countries. It has proven its resilience by emerging relatively unscathed from acute financial and fiscal crises in the region.

However, the most difficult and challenging steps towards the economic integration envisaged in the Treaty of Asunción remain to be taken.

See Also

- ▶ Currency Unions
- ▶ Foreign Direct Investment
- ▶ Regional and Preferential Trade Agreements

Bibliography

Berlinski, J. 2004. *Los impactos de la política comercial: Argentina y Brazil (1988–1997)*. Buenos Aires: Siglo XXI.

Bohara, A., K. Gawande, and P. Sanguinetti. 2004. Trade diversion and declining tariffs: Evidence from Mercosur. *Journal of International Economics* 64: 65–88.

Bouzas, R. 2002. Mercosur after ten years: Learning process or déjà-vu? In *Paths to regional integration: The case of Mercosur*, ed. J. Tulchin and R. Espach. Washington, DC: Woodrow Wilson International Center.

Brambilla, I. 2005. A customs union with multinational firms: The automobile market in Argentina and Brazil. Working paper no. 11745. Cambridge: NBER.

Cadot, O., J. de Melo, and M. Olarreaga. 2003. The protectionist bias of duty drawbacks: Evidence from Mercosur. *Journal of International Economics* 59: 161–182.

Castilho, M., and S. Zignago. 2002. Trade effects of FDI in Mercosur: A disaggregated analysis. In *An integrated approach to the EU-Mercosur association*, ed. P. Giordano. Paris: Sciences Po.

Chang, W., and L. Winters. 2002. How regional blocs affect excluded countries: The price effects of Mercosur. *American Economic Review* 92: 889–904.

Chudnovski, D. 2001. *El boom de la inversión extranjera directa en el MERCOSUR*. Buenos Aires: Siglo XXI.
Eichengreen, B. 1998. Does Mercosur need a single currency? Working paper no. 1018, Center for International and Development Economic Research, UC Berkeley.
Olarreaga, M., and I. Soloaga. 1998. Endogenous tariff formation: The case of Mercosur. *World Bank Economic Review* 12(1): 297–320.
Porto, G. 2006. Using survey data to assess the distributional effects of trade policy. *Journal of International Economics* 70(1): 140–160.
Schiff, M., and W. Chang. 2003. Market presence, contestability and the terms-of-trade effects of regional integration. *Journal of International Economics* 60: 161–175.
Vaillant, M. 2005. Mercosur: Southern integration under construction. *Internationale Politik und Gesellschaft* 2: 52–71.
Yeats, A. 1998. Does Mercosur's trade performance raise concerns about the effects of regional trade arrangements? *World Bank Economic Review* 12(3): 1–28.

Merger Analysis (United States)

Dennis W. Carlton and Jeffrey M. Perloff

Abstract

There are three different types of mergers: horizontal, vertical, and conglomerate. We discuss all three and explain why mergers can be a desirable way to expand a firm. Then we turn to the evidence on the amount of merger activity. Finally, we address one of the important questions surrounding mergers: whether they are motivated by the desire to improve efficiency or by the desire to acquire market power. Although the evidence is sometimes ambiguous, the overwhelming consensus is that most merger activity in the United States is motivated by efficiency considerations.

Keywords

Antitrust enforcement; Conglomerate mergers; Economies of scale; Economies of scope; Efficiency; Horizontal mergers; Litigation; Market power; Mergers; Reputation; Takeovers; Taxation of corporate profits; Transaction costs; Vertical mergers

JEL Classifications

L40

In the economics literature, a merger is the combination of the assets of two or more firms. Economists usually distinguish three different types of merger: horizontal, vertical, and conglomerate. Horizontal mergers are between rivals; vertical mergers involve firms one of which supplies inputs to the other(s); conglomerate mergers are between firms in unrelated businesses. Mergers represent one way for a firm to acquire assets as an already assembled package.

We first discuss why a merger is sometimes a desirable way to expand a firm. Then we turn to the evidence on the amount of merger activity. Finally, we address one of the important questions surrounding mergers: whether they are motivated by the desire to improve efficiency or by the desire to acquire market power. Although the evidence is sometimes ambiguous, the overwhelming consensus is that most merger activity in the United States is motivated by efficiency considerations.

Reasons for Mergers

The most important obvious reason for mergers is to increase efficiency. There is a variety of ways in which mergers can enhance efficiency. By increasing its size, a firm may be able to achieve economies of scale in production, distribution, management, or other aspects of the firm's operation, such as research and development. By eliminating duplication of certain management functions, firms may be able to cut their total costs. Certain scale efficiencies may arise naturally when firms are regulated or have reporting requirements. For example, a merged firm may have to submit tax and other government forms only once as a result of the merger.

By increasing the number of its activities, the merged firm may achieve economies of scope, efficiencies that result from engaging in related activities done together in one firm. For example, the ability of one firm to provide a wide range of products may make distribution easier. Alternatively, the ability to use in one activity knowledge

gained in another can make it more efficient to have one firm perform both activities rather than having each activity performed by a different firm.

A common reason for vertical mergers is to eliminate transaction costs associated with using the marketplace to obtain supplies. (Of course, there is the offsetting cost of running a larger firm.) An example of a transaction cost is opportunism. In marketplace transactions, a buyer may (unexpectedly) be able to exploit the seller (or vice versa). For example, the seller may have no other possible buyer in the short run, and the buyer could demand a lower price than the once originally agreed upon. A vertical merger is an alternative to other mechanisms such as reputations or contract litigation to deal with this problem.

A vertical merger may eliminate the distortion of an upstream (input) monopoly. Prior to merger, the downstream (output) firm decides how to produce and how to price its output based on this distorted input price. If the output is produced with variable input proportions, there is a loss of efficiency to the economy that a vertical merger can fix. There is a private incentive to vertically integrate, but the effect on social welfare is ambiguous.

If both the upstream and the downstream firms are non-competitive, a vertical merger eliminates 'double marginalization'. An upstream firm with market power raises its price above its marginal cost. Then the downstream firm adds an additional markup so that the final consumers pay a double markup. If the firms merge, they set only a single markup, causing output price to fall, output to expand and social welfare to rise.

So far, these explanations do not answer the question why one firm merges with another rather than buying the underlying assets and assembling them itself. Aside from competitive effects, one answer is that another firm is a package of already assembled assets and it may be cheaper to buy an existing firm than to create one.

Mergers can also be used to transfer assets from the control of bad managers (or investors) to good ones. Suppose that Firm X has very smart managers, while Firm Y has either incompetent managers or managers who are not performing well because no one is monitoring their actions. Here, a transfer of assets to X should allow Y's assets to be more productive. X should be able to pay more for Y's assets than they are worth based on the market's valuation of Y's cash flows under its current incompetent management. This disparity in value creates an incentive for X to purchase Y. To avoid being taken over, Y's managers can improve their performance (that is, the takeover threat disciplines them) or engage in defensive tactics designed to thwart such a takeover in order to save their jobs. If these defensive tactics induce the acquiring firm to raise its price for Y, the tactics can benefit Y's shareholders. There is a large literature on defensive tactics as well as their sometimes ambiguous efficiency consequences. In a hostile takeover, X buys Y despite the desire of Y (or its managers) to remain independent. The use of hostile takeovers in the 1980s coincided with the ability of acquiring firms to obtain financing through junk bonds (bonds below investment grade).

Aside from efficiency motivations, another rationale for a merger is to eliminate competition between the merging firms. The antitrust laws of the United States forbid mergers that result in a lessening of competition with a consequent increase in price. Although antitrust concerns about mergers mainly arise in the context of horizontal mergers, such concerns can also arise with vertical mergers. One concern is that a vertical merger could eliminate a key supplier for a rival firm. Typically, theories of vertical harm are much less certain in their predictions than theories of competitive harm arising from horizontal mergers.

In addition to efficiency and market power explanations, there is a variety of other reasons for mergers. Tax considerations can sometimes make it advantageous for one firm to merge with another. For example, if one firm has a loss and another a profit, a merger can lower their total tax liability. The merged firm may be able to report no profit and therefore owe no corporate tax. Separately, one of the firms (the profitable one) would have to pay a tax. Mergers can also allow managers to engage in empire building, or allow a firm to have an 'excuse' ('I'm no longer in charge') to renege on certain informal promises made to workers or other firms.

Evidence: Merger Activity

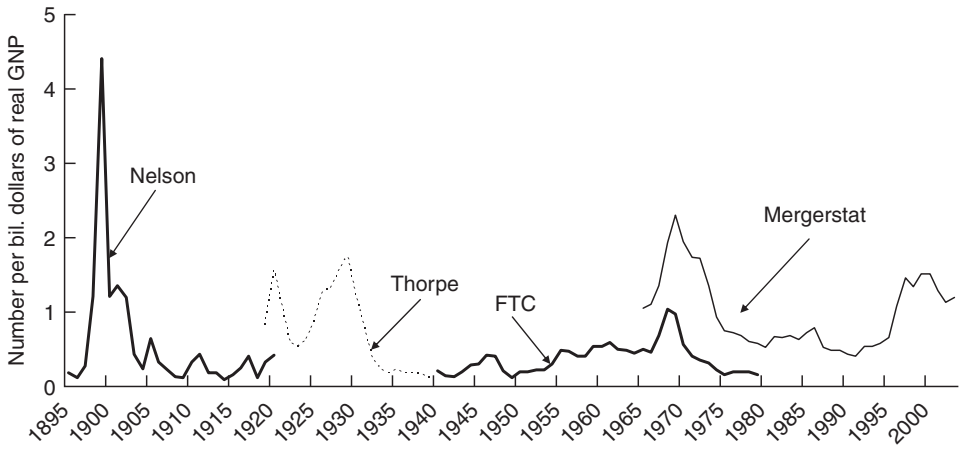
Mergers come in waves, being common in certain industries at certain times. Because no single data series on merger activity goes back to 1900, we must splice together sometimes inconsistent data sources to study mergers over the 20th century. Figure 1 presents data on the amount of US merger activity relative to the size of the economy back to 1900. By controlling for the economy's size (a larger economy is likely to generate more merger activity), we can compare the intensity of merger activity at different times.

Figure 1 indicates that there have been several waves of merger activity. The first, around 1900, was (relatively) the largest and represents the creation of some of the best-known firms in the United States, such as General Electric and U.S. Steel. This was a time of great change with significant developments in transportation and communications. The second wave was in the 1920s and helped to create several oligopolies. The third, in the 1960s, involved conglomerate mergers. In the 1980s, the fourth wave (which would be more evident in the figure if we had dollar value of merger activity instead of the number of mergers) arose as hostile takeovers became popular in the United States. The fifth was in the late 1990s and disproportionately involved airlines, telecommunications, banking and other industries that had previously been heavily regulated.

The timing of merger waves seems to coincide with stock market booms for reasons no one has completely explained. One recent explanation by Shleifer and Vishny (2003) maintains that during stock booms stocks are overvalued (a fact known by managers but not outside investors) and managers use the (overvalued) stock to purchase other firms. In stock market booms, the use of stock rather than cash to buy other firms' assets does increase, consistent with the hypothesis.

Empirical Evidence: Rationales for Mergers

A central question is whether mergers improve efficiency or, especially in the case of horizontal mergers, reduce competition and harm consumers. Despite an enormous number of studies of this question, the answer is still somewhat controversial. Our conclusion is that, although there is no doubt that some mergers are poorly motivated and turn out badly for the firms, and that some



Merger Analysis (United States), Fig. 1 Annual number of mergers and acquisitions per billion US dollars of real GNP (United States, 1895–2003). *Note:* 1982 dollars (*Sources:* Nelson Series, Federal Trade Commission (FTC) 'Board' series and Mergerstat). In 2003 the Bureau of Economic Analysis made comprehensive revisions to its National Income and Product Accounts. These revised figures were used to calculate deflators for the FTC series and Mergerstat. The Mergerstat series has broader coverage (for example, more industries, lower thresholds for reporting) than the FTC series (Adapted from Golbe and White (1988), Figure 9.7)

horizontal mergers reduce competition and harm consumers, most are expected to be profitable, to enhance efficiency and not to reduce competition.

Researchers have used three types of data: stock market data, accounting data, and price or output data. Because of their availability, stock market data have been used most often. Stock market studies rely on the premise that stock market prices are a good indication of a firm's expected future profitability (and make subtle assumptions about when information gets reflected in prices). These stock market studies can capture the effect of a merger on the acquiring firm, the acquired (target) firm, and rivals. Accounting data may have certain biases that can be hard to correct, and can be difficult to obtain. The same is true of data on price and output. In contrast to stock market studies, studies using either accounting or performance data are *ex post* studies of mergers (what happened after the mergers), while studies using stock market data are *ex ante* studies (what is expected to happen). We present a brief summary of the major findings (see Carlton and Perloff 2005, and the references cited there, especially Andrade and Stafford 2001, and Pautler 2003).

Shareholders of an acquired firm earn a premium of between 16 and 25 per cent above the price prevailing prior to the merger. This premium is now higher than it used to be before the Williams Act of 1968 was passed. The Williams Act requires a firm to reveal publicly its intentions to acquire another firm.

Shareholders of the acquiring firm do not do very well. Although they earned slightly positive returns in the 1960s (plus four per cent), their returns became slightly negative (minus three per cent) in the 1980s and 1990s. Interestingly, the form of the acquisition (whether cash or stock) influences the return, with acquirers doing better when more cash is used, though it is unclear why this should occur. (The use of stock to finance mergers has increased over time, with about 60 per cent of transactions in the 1990s financed entirely by stock.)

Overall, the total return (which is what matters for efficiency) to the combined acquiring and acquired firms is positive. That is, the total value of the merged firm is about 2–7.5 per cent higher after a merger than the sum of each firm's value pre-merger.

Researchers using accounting or other performance data have had more difficulty documenting gains from mergers. Using data from the 1960s and 1970s, Scherer (1988) and Ravenscreft and Scherer (1987) do not find increased profits post-merger. Andrade and Stafford (2001) use Scherer's data and show that the data support the efficiency hypothesis if one controls for industry benchmarks. Lichtenberg and Siegal (1987) find significant positive effects of mergers on productivity.

Studies of stock markets and of individual industries have been used to investigate whether horizontal mergers generally create market power. The stock market studies exploit the idea that a merger that creates efficiency will cause the stock price of the (to-be-merged) firm to rise and that of its rivals to fall. In contrast, a horizontal merger that eliminates a rival should be expected to also benefit other rivals (since industry price will rise if competition is eliminated). Banerjee and Eckard (1998) show that, even for the massive merger wave around 1900 (prior to strict enforcement of antitrust laws forbidding mergers that eliminated competition), rival firms suffered as a result of a horizontal merger, supporting the efficiency hypothesis. There are of course exceptions, and some studies of recent mergers (for example, some airline mergers) show that horizontal mergers can harm competition and raise price. However, most of the literature (though certainly not all) supports the view that mergers generally should be expected to help consumers.

See Also

- ▶ Firm Boundaries (Empirical Studies)
- ▶ Merger Simulations
- ▶ Mergers, Endogenous

Bibliography

Andrade, G., and E. Stafford. 2001. New evidence and perspectives on mergers. *Journal of Economic Perspectives* 15(2): 103–120.

This article draws heavily on Carlton and Perloff (2005, ch. 2). The reader interested in more detailed discussion should consult that work together with the references cited therein.

Banerjee, A., and E. Eckard. 1998. Are mega-mergers anti-competitive? Evidence from the first great merger wave. *RAND Journal of Economics* 29: 803–827.
Carlton, D., and J. Perloff. 2005. *Modern industrial organization*. New York: Pearson Addison Wesley.
Golbe, D., and L. White. 1988. A time series analysis of mergers and acquisitions in the US Economy. In *Corporate takeovers: Causes and consequences*, ed. A. Auerbach. Chicago: University of Chicago Press.
Lichtenberg, F., and D. Siegal. 1987. Productivity and changes in ownership of manufacturing plants. *Brookings Papers on Economic Activity* 1987(3): 63–83.
Pautler, P. 2003. Evidence on mergers and acquisitions. *Antitrust Bulletin* 48: 119–221.
Ravenscraft, D., and F. Scherer. 1987. *Mergers, sell-offs and economic efficiency*. Washington, DC: Brookings Institution.
Scherer, F. 1988. Corporate takeovers: The efficiency arguments. *Journal of Economic Perspectives* 2(1): 69–82.
Shleifer, A., and R. Vishny. 2003. Stock market driven acquisitions. *Journal of Financial Economics* 70: 295–311.

Merger Simulations

Aviv Nevo

Keywords
Airline industry; Merger simulations; Mergers

JEL Classification
L13

The key in an evaluation of a proposed merger is to determine whether the reduction of competition it would cause is outweighed by potential cost reductions. Traditional analysis of mergers is primarily based on industry-concentration measures. A market is defined and market shares of the relevant firms are used to compute a pre-merger concentration measure as well as a change in this measure due to the merger. Both the pre-merger level and the change in concentration are then compared with preset levels. The intuition is that, if the industry is concentrated, or if the change in concentration is large, then the anti-competitive effect will dominate. Using this approach to evaluate mergers in some industries is problematic for at least two reasons. In many cases the product offerings make the definition of the relevant product (or geographic) market difficult. Even if the relevant market can be defined, the computed concentration index provides a reasonable standard by which to judge the competitive effects of the merger only under strong assumptions.

Merger simulation attempts to deal with these challenges. The basic idea consists of 'front-end' estimation, in which the structural primitives of the model are estimated, and a 'back-end' analysis, in which the estimates are used to simulate the post-merger equilibrium. The approach proceeds as follows.

First, demand parameters are recovered by econometric estimation, if the data are rich enough, or, if data (with enough variation) are not available, then marketing and other anecdotal evidence can be used to approximate the effects of prices on demand (Werden and Froeb 1994). Estimation has to deal with two main challenges: a flexible functional form, especially with a large number of products, and reasonable identifying assumptions. The most commonly used approaches, to deal with the large number of products, are multi-level budgeting (Hausman et al. 1994) and the discrete-choice, characteristics, approach (Berry et al. 1995; Nevo 2000). Prices are set endogenously and typically respond to demand shocks that are unobserved by the researcher, and therefore instrumental variables are needed. Two common instrumental variables are observed characteristics of other products (Bresnahan 1987; Berry et al. 1995) and out-of-market prices (Hausman et al. 1994; Nevo 2000).

Second, pre-merger cost parameters are recovered. One approach is to assume a model of pricing (Bertrand, say) and to use it jointly with the estimated demand parameters to recover implied marginal costs. If needed, the implied marginal costs can be regressed on characteristics in order to recover cost functions. Alternatively, the pricing equation, and the cost functions, can be estimated jointly with demand. Either way, the model of pricing can, and should, be tested (Porter 1983; Bresnahan 1987; Nevo 2001). Finally, marginal cost can be

approximated from accounting data, but these tend to be unreliable.

Third, the recovered marginal costs and estimated demand parameters are used jointly to simulate the new equilibria that would result from a merger. Usually, the analysis focuses on 'unilateral effects', with the likelihood of (tacit) collusion fixed. In principle, however, the simulation can use a different model of competition post-merger from the one used to recover the parameters. In order to address potential cost reductions, the simulation can be performed with marginal cost fixed, by changing marginal costs or by asking what cost saving is required to keep consumer welfare, or any other measure, at a certain level (Nevo 2000). Finally, the model can be used to assess the likelihood of entry and/or the change in incentive to collude.

The end result is a prediction of post-merger prices and quantities under several scenarios. With the use of the estimated demand and supply functions, these equilibrium quantities can be converted into consumer welfare and (variable) profits. The change in welfare and profits can be used as the basis for evaluating the merger instead of the change in concentration. This has the advantage of being linked to economic theory and the underlying trade-off between reduction in competition and improved efficiency. It also allows the parties to assess the accuracy of the prediction due to the assumptions by simulating under different assumptions, or due to the data by computing standard errors.

There are several potential pitfalls in using merger simulation. The simulation is only as good as the model it is based on and the parameter estimates that go into the simulation. Therefore, one should take extra care in choosing a model suitable for the industry. Furthermore, in some cases data and time constraints might limit the ability to consistently estimate the parameters required for the simulation.

Despite the fact that merger simulation has been used extensively in practice, there is little work testing its accuracy with the use of post-merger data. One exception is a study of mergers in the airline industry (Peters 2003) that finds that simulation methods do a reasonable job at predicting the price effects of mergers. Peters also finds that a large fraction of the unexplained change in prices comes from changes in marginal costs or firm conduct (his analysis cannot separate the two). Retrospective analysis of this sort is useful not just in evaluating the quality of predictions but also in pointing to directions in which the modelling and analysis can be improved.

For further readings and details see Whinston (2005, ch. 3).

Bibliography

Berry, S., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890.

Bresnahan, T. 1987. Competition and collusion in the American automobile oligopoly: The 1955 price war. *Journal of Industrial Economics* 35: 457–482.

Hausman, J., G. Leonard, and J. Zona. 1994. Competitive analysis with differentiated products. *Annales d'Economie et de Statistique* 34: 159–180.

Nevo, A. 2000. Mergers with differentiated products: The case of the ready-to-eat cereal industry. *RAND Journal of Economics* 31: 395–421. Reprinted in *Empirical industrial organization*, ed. P. Joskow and M. Waterson. Cheltenham: Edward Elgar, 2004.

Nevo, A. 2001. Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69: 307–342.

Peters, C. 2003. *Evaluating the performance of merger simulation: Evidence from the US airline industry*. Working paper No. 32. Center for the Study of Industrial Organization, Northwestern University.

Porter, R. 1983. A study of cartel stability: The Joint Executive Committee, 1880–1886. *Bell Journal of Economics* 14: 301–314.

Werden, G., and L. Froeb. 1994. The effects of mergers in differentiated products industries: Logit demand and merger policy. *Journal of Law, Economics, and Organization* 10: 407–426.

Whinston, M. 2005. *Lectures on antitrust economics*. Cambridge, MA: MIT Press.

Mergers

G. Meeks

In any one year it is not uncommon in the US or the UK for firms representing one per cent of the

assets of the company sector to be acquired by others in a merger or takeover (in economic, if not in legal, terms the two are often indistinguishable). In one of the cyclical peaks in merger activity the figure has risen to almost three per cent in the US (Federal Trade Commission 1977; Scherer 1980) and five per cent in the UK (Singh 1975). For the acquiring firms, growth by merger can be a very significant form of expansion: to take an extreme example, UK listed companies in aggregate spent more in 1968 on acquiring second-hand assets through merger than they did on new fixed investment (Meeks 1977).

Three categories of merger are commonly distinguished when the motives for merger and its consequences are being analysed: horizontal merger, between competitors; vertical merger, between supplier and customer; and conglomerate merger, between companies with no complementary markets or production processes. In practice, individual mergers often refuse to fit neatly into just one of these categories and allocating aggregate merger activity to them involves some, often arbitrary, assumptions. Despite this qualification, two generalizations are warranted. First, the share of conglomerate merger in total merger activity has been rising on both sides of the Atlantic. But secondly, the share is much higher in the United States than in the UK or in most other Western economies: by the 1960s and 1970s conglomerate merger accounts for around 80 per cent of the total value of mergers in the US (Scherer 1980) but for only a minority of merger activity in the UK (Goudie and Meeks 1982).

This difference between the two countries is normally attributed to the contrasting merger policies of their governments for most of the postwar period. In the US most sizeable horizontal mergers that might have been proposed would have been outlawed, whereas in the UK no government constraint at all was imposed on merger until 1965 and subsequent policy is generally regarded (e.g. by O'Brien 1978; Meeks and Meeks 1981) as very permissive by American standards. The more restrictive United States policy was designed to maintain competition: US merger waves in earlier periods had led to rapid concentration of sales in many markets (Nelson 1959; Stigler 1950). And in the absence of tight controls, merger activity in the UK has raised seller concentration significantly: one study which analyses the shares of different industries' sales supplied by the top five firms assigns most of the substantial rise in these shares between 1957 and 1969 to the merger process (Hannah 1983). Even where full monopoly is not achieved through merger it is held that concentration through merger can often lead to comparable results under oligopoly: as the number of dominant firms declines the potential is increased for collusive behaviour in the interests of firms' owners or managers and at the expense of the customer (Hannah and Kay 1977).

But the monopolization motive for horizontal merger is seen by some as quite outweighed by another motive: cost reduction. Achieving scale economies has been seen as an objective which, in contrast with monopolization, reconciles private and social interests. These economies may arise at the plant level (Pratten 1971) or as a result of operating several plants within one firm (Scherer et al. 1975): in either case horizontal merger may bring together firms which, individually, fall short of the minimum efficient scale.

Whatever the potential gains from scale, however, there is in practice little evidence of wide spread cost reduction following merger. A range of studies with diverse methods has failed to find cost improvements following the majority of mergers (Newbould 1970; Singh 1971; Utton 1974; Meeks 1977; Cowling et al. 1980, Mueller (ed.), 1981; Kumar 1984). And with scarcely an exception even the studies which focus on profitability (which might be expected to benefit doubly from horizontal merger – through both the enhancement of monopoly power and the realization of scale economies) find no improvement following merger for the average firm.

Another motive for merger but this time peculiar to conglomerate merger is diversification, reducing the firm's dependence on its existing line(s) of business. Combining activities with uncorrelated (or, better still, negatively correlated) returns produces lower variability for the group's profit rate (the smaller variability of larger, more diversified firms is well documented: see, e.g., Samuels and Smyth (1968)). And this lower

variability is held to reduce the merged firms' cost of capital. In a perfect capital market with no transactions costs, however, any gains to the share price from combining earnings streams which are not positively correlated should have been exhausted by the efforts of investors to achieve 'home-made' diversification of their portfolios. In principle, then, such gains could be realized by shareholders in the absence of merger: if merger is required market imperfections are the cause.

A number of other motives for merger also hinge on market imperfections but are not limited to conglomerate merger. The very imperfect information available to investors has been suggested as one reason for the relatively favoured capital market treatment accorded to larger, more prominent companies compared with smaller, less well known ones. Certainly the evidence is that larger companies are accorded a higher price-earnings ratio than smaller ones and that this favour is not explained away by factors such as the larger firms' smaller earnings variability (see Prais 1976). This discrimination in favour of large firms means that the earnings of a small company are more highly valued on the stock market if it becomes a subsidiary of a giant company than if it remains independent; and if the discrepancy in price-earnings ratios is significant it may pay the shareholders to support merger even if costs rise as a result and the subsidiary's earnings actually fall (Lynch 1971).

The tax system too sometimes offers firms' owners incentives to take part in merger. To take a British example, given the disparity between tax rates on income and those on capital gains, a proprietor may gain in post-tax terms from selling his business and realizing relatively lightly taxed capital gains even if the merger causes some loss in profitability for the business (Hannah and Kay 1977). Then again, merger may ease the transition to a capital structure which secures more favourable tax treatment, a transition which tax regulations may forbid in the absence of merger (Department of Trade and Industry 1978).

In those companies where control is divorced from ownership, and where the firms' salaried managers enjoy considerable discretion over their strategy, the motives of managers rather than of owners will be more significant in explaining merger. And merger (horizontal, vertical or conglomerate) may enhance managers' income and power. There is a strong association between the growth of the firm (whether by merger or by 'internal' expansion) and the growth of its senior managers' income (Cosh 1975; McEachern 1975; Meeks and Whittington 1975). And the aggregate concentration of power in the economy is well documented too: in the US, mergers have helped the largest 100 corporations to raise their share of manufacturing valued added from some 23 per cent in 1947 to 33 per cent by the early 1970s (Scherer 1980); whilst the UK, with its more permissive government stance on merger, has seen the corresponding share rise from 22 per cent in 1949 to 41 per cent in 1970 (Prais 1976). Of course, merger is not the only means of achieving growth; but it often permits unusually rapid rates of expansion since it mitigates 'Penrose effects' – the difficulties of assimilating numerous additions to the management team (Penrose 1959): whole subsidiary management teams are acquired intact.

Ironically, the constraint on rapid expansion which has been emphasized by growth theorists of the firm such as Marris (1964) is the growing firm's fear of being itself taken over: if the owners' interest in profit is sacrificed to the executives' pursuit of growth, it is held that the company's share price will fall, reflecting the shortfall of actual profit performance below potential. And the further the share price falls, the greater is the incentive for a profit-oriented raider to acquire the errant firm and restore its profitability and market valuation.

Such corrective takeovers have been seen by some as providing a crucial control mechanism in modern capitalist economies. It is held that even if product markets are imperfect and leave managers some discretion over their business objectives, and even if there is a widespread divorce of ownership from control, still the market for company control will ensure profit maximization: errant firms will be prey to hostile merger (see e.g. Manne 1965; Meade 1968).

But the potential effectiveness of this mechanism has been called into question by economic theorists. For example, Alchian (1950) and Winter (1964) show that in certain realistic conditions

economic natural selection does not ensure the survival of profit maximizers; whilst Grossman and Hart (1980) explore the role of 'free-riding' shareholders in inhibiting 'disciplinary' takeovers. And the empirical evidence available proves inimical to the notion that fear of takeover acts as a potent deterrent to inefficiency. In the UK at least, not only do large numbers of relatively unprofitable firms survive for long periods without being taken over (Whittington 1971), whilst the typical merger victim seems to have achieved about average profitability (Singh 1971; Meeks 1977) and profitability is not typically enhanced by merger; but also survival statistics suggest that increased size rather than higher profitability appears more likely to secure immunity from takeover for the large unprofitable firm (Singh 1975). Perversely, then, the takeover threat seems in practice to provide not so much a stimulus to efficiency as an incentive to embark on yet more takeovers; for merging with other companies, even in the absence of efficiency gains, is one means of finding shelter from being taken over oneself.

See Also

- ▶ Cartels
- ▶ Conglomerates
- ▶ Corporate Economy
- ▶ Industrial Organization
- ▶ Market Structure
- ▶ Rationalization of Industry
- ▶ Vertical Integration

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58: 211–221.

Cosh, A.D. 1975. The remuneration of chief executives in the United Kingdom. *Economic Journal* 85: 75–94.

Cowling, K., et al. 1980. *Mergers and economic performance*. Cambridge: Cambridge University Press.

Department of Trade and Industry. 1978. *A review of monopolies and merger policy: A consultative document*. London: HMSO.

Federal Trade Commission. 1977. *Statistical report on mergers and acquisitions*. Washington, DC: Government Printing Office.

Goudie, A.W., and G. Meeks. 1982. Diversification by merger. *Economica* 49(196): 447–459.

Grossman, S., and O.D. Hart. 1980. Takeover bids, the free-rider problem, and the theory of the corporation. *Bell Journal of Economics* 11(1): 42–64.

Hannah, L. 1983. *The rise of the corporate economy*. London: Methuen.

Hannah, L., and J. Kay. 1977. *Concentration in modern industry*. London: Macmillan.

Kumar, M.S. 1984. *Growth, acquisition and investment*. Cambridge: Cambridge University Press.

Lynch, H.H. 1971. *Financial performance of conglomerates*. Cambridge, MA: Harvard University Press.

Manne, H.G. 1965. Mergers and the market for corporate control. *Journal of Political Economy* 73: 110–120.

Marris, R. 1964. *The economic theory of 'Managerial' capitalism*. London: Macmillan.

McEachern, W.A. 1975. *Managerial control and performance*. Lexington: Health.

Meade, J.E. 1968. Is 'the new industrial state' inevitable? *Economic Journal* 78: 372–392.

Meeks, G. 1977. *Disappointing marriage: A study of the gains from merger*. Cambridge: Cambridge University Press.

Meeks, G., and J.G. Meeks. 1981. The case for a tighter merger policy. *Fiscal Studies* 2(2): 33–46.

Meeks, G., and G. Whittington. 1975. Directors' pay, growth and profitability. *Journal of Industrial Economics* 24(1): 1–14.

Mueller, D.C. (ed.). 1981. *The determinants and effects of merger: An international comparison*. Cambridge, MA: Oelgeschlager\Gunn and Hain.

Nelson, R.L. 1959. *Merger movements in American industry, 1895–1956*. Princeton: Princeton University Press.

Newbould, G.D. 1970. *Management and merger activity*. Liverpool: Guthstead.

O'Brien, D. 1978. Mergers – time to turn the tide. *Lloyds Bank Review* (130): 32–44.

Penrose, E.T. 1959. *The theory of the growth of the firm*. Oxford: Basil Blackwell.

Prais, S.J. 1976. *The evolution of giant firms in Britain*. Cambridge: Cambridge University Press.

Pratten, C.F. 1971. *Economies of scale in manufacturing industry*. Cambridge: Cambridge University Press.

Samuels, J.M., and D.J. Smyth. 1968. Profits, variability of profits and firm size. *Economica* 35: 127–139.

Scherer, F.M. 1980. *Industrial market structure and economic performance*. Chicago: Rand McNally.

Scherer, F.M., A. Beckensten, E. Kaufer, and R.D. Murphy. 1975. *The economics of multi-plant operation: An international comparisons study*. Cambridge, MA: Harvard University Press.

Singh, A. 1971. *Takeovers*. Cambridge: Cambridge University Press.

Singh, A. 1975. Takeovers, economic natural selection and the theory of the firm: Evidence from the postwar United Kingdom experience. *Economic Journal* 85: 497–515.

Stigler, G.J. 1950. Monopoly and oligopoly by merger. *American Economic Review: Papers and Proceedings* 40: 23–34.
Utton, M.A. 1974. On measuring the effects of industrial mergers. *Scottish Journal of Political Economy* 21(1): 13–28.
Whittington, G. 1971. *The prediction of profitability and other studies of company behaviour*. Cambridge: Cambridge University Press.
Winter Jr., S.G. 1964. Economic 'natural selection' and the theory of the firm. *Yale Economic Essays* 4(1): 225–272.

Mergers, Endogenous

Volker Nocke

Keywords

Antitrust; Bidding games; Coalition formation games; Collusion; Concentration; Cournot model; Endogenous mergers; Exogenous mergers; Horizontal mergers; Market power; Mergers; Monopolization; Oligopoly; Vertical mergers

JEL Classifications

L40

The term 'endogenous mergers' reflects the view in economic theory that mergers are equilibrium outcomes. The literature on endogenous mergers explicitly analyses firms' incentives to merge and makes predictions on the volume and type of mergers that are likely to occur. In this literature, merger formation is modelled as a bidding game or non-cooperative coalition formation game (Kamien and Zang 1990; Gowrisankaran 1999; Nocke 2000; Pesendorfer 2005), or as an anonymous merger market where firms can buy or sell corporate assets (Jovanovic and Rousseau 2002; Nocke and Yeaple 2007). The literature on endogenous mergers is conceptually distinct from the literature on *exogenous* mergers, which considers the positive and normative effects of a merger between a given ('exogenous') set of firms.

To analyse the endogenous merger process, one first needs to understand why firms may want to merge. Several motives for mergers have been identified in the literature.

First, firms may want to merge to realize efficiency gains or 'synergies'. Mergers may allow firms to exploit complementarities in their capabilities (Nocke and Yeaple 2007), or they may be an efficient way to reallocate used capital from less productive firms to more productive firms (Jovanovic and Rousseau 2002).

Second, firms may want to merge to increase their market power. However, as Salant et al. (1983) have shown for the Cournot model, a merger solely aimed at increasing market power may not be profitable: to the extent that merging firms want to reduce joint output to raise price, non-participating outsiders will increase their output in response, imposing a negative externality on the merging firms. (This point relies heavily on the Cournot assumption; see Deneckere and Davidson 1985.) While it has generally been acknowledged that horizontal mergers (between firms competing in the same market) may lead to higher prices and lower welfare, the Chicago School of antitrust has long held the view that vertical mergers (between upstream suppliers and their downstream customers) are efficiency-enhancing. By showing that vertical mergers may allow foreclosure of upstream suppliers or downstream buyers, this view has recently been refuted in a series of articles (see Rey and Tirole 2005, for a survey).

Third, firms may want to merge to facilitate collusion. A horizontal merger may facilitate collusion by reducing the number of players in the industry, or by reallocating industry capacity in a way that equalizes firms' incentives to cheat (Compte et al. 2002). A vertical merger may facilitate upstream collusion by reducing the number of downstream outlets through which an upstream firm can profitably deviate. Furthermore, to the extent that collusion is sustainable only if the vertically integrated firm receives a larger market share than an unintegrated firm, firms may have an incentive to merge so as to demand and obtain a larger share of the collusive pie (Nocke and White 2003).

Finally, a variety of other motives for merger have been proposed, some of which are based on the view that firms do not necessarily maximize profits. For example, it has been argued that

managers may have an incentive to engage in empire building.

Focusing on the market power motive, much of the recent literature on endogenous mergers has been concerned with studying the limits to monopolization through mergers and acquisitions, and making predictions on the relationship between concentration levels and industry characteristics (Kamien and Zang 1990; Nocke 2000; Gowrisankaran and Holmes 2004). The starting point of this literature is the observation by Stigler (1950, pp. 25–6) that 'the promoter of a merger is likely to receive much encouragement from each firm – almost every encouragement, in fact, except participation'.

To understand Stigler's point that a merger to monopoly may not obtain even when feasible, consider an industry with N firms, each running a single plant to produce a homogeneous or differentiated good. If a subset of these firms merge, they will internalize any externality in the price/output decisions they impose on each other. Unless efficiency gains from merging are large, a merged entity would thus produce a smaller output per plant than a single-plant firm: a firm participating in the merger ('insider') would be better off than a firm not participating ('outsider'). Let $\Pi(N; 0)$ denote monopoly profits, and $\Pi(1; N-1)$ the profit of a single-plant firm competing with a larger firm owning $N-1$ plants. Assume the merger would take place even when only $N-1$ firms agreed to merge. Then, firm i will agree to merge with its $N-1$ rivals only if $\Pi(1; N-1) \leq s_i \Pi(N; 0)$, where s_i is firm i's equity share in the merged entity. Since this must hold for any firm i, merger to monopoly will occur only if $\Pi(1; N-1) \leq \Pi(N; 0)/N$. In standard oligopoly models, this inequality is often violated if efficiency gains from merging are small, the number of firms is large, and competition is not too 'tough'. Merger to monopoly may thus fail to occur, even though it would maximize joint profits, as some firm(s) may be better off staying outside and taking a free ride on the merged entity's effort to restrict output.

There may also be limits to monopolization through mergers and acquisitions because of entry. To the extent that a merger makes the industry *less* competitive, a merger between incumbents may induce more entry in the future, reducing the incumbents' profits. By not merging with their rivals, incumbent firms may thus credibly commit to compete vigorously and deter further entry.

See Also

- ▶ Cartels
- ▶ Merger Analysis (United States)
- ▶ Merger Simulations

Bibliography

Compte, O., F. Jenny, and P. Rey. 2002. Capacity constraints, mergers and collusion. *European Economic Review* 46: 1–29.

Deneckere, R., and C. Davidson. 1985. Incentives to form coalitions with Bertrand competition. *RAND Journal of Economics* 16: 473–486.

Gowrisankaran, G. 1999. A dynamic model of endogenous horizontal mergers. *RAND Journal of Economics* 30: 56–83.

Gowrisankaran, G., and T.J. Holmes. 2004. Mergers and the evolution of industry concentration: Results from the dominant firm model. *RAND Journal of Economics* 35: 561–582.

Jovanovic, B., and P.L. Rousseau. 2002. The Q-theory of mergers. *American Economic, Review Papers and Proceedings* 92: 198–204.

Kamien, M.I., and I. Zang. 1990. The limits of monopolization through acquisition. *Quarterly Journal of Economics* 105: 465–499.

Nocke, V. 2000. Monopolisation and industry structure. Economics Working Paper No. 2000-W27, Nuffield College, Oxford.

Nocke, V., and L. White. 2003. Do vertical mergers facilitate upstream collusion? Working Paper No. 03-033, PIER, University of Pennsylvania.

Nocke, V., and S. Yeaple. 2007. Cross-border mergers and acquisitions versus greenfield foreign direct investment: The role of firm heterogeneity. *Journal of International Economics* 72(2): 336–365.

Pesendorfer, M. 2005. Mergers under entry. *RAND Journal of Economics* 36: 661–679.

Rey, P., and J. Tirole. 2005. A primer on foreclosure. In *Handbook of industrial organization*, ed. M. Armstrong and R. Porter, Vol. 3. Amsterdam: North-Holland.

Salant, S.W., S. Switzer, and R.J. Reynolds. 1983. Losses from horizontal merger: The effects of an exogenous change in industry structure on Cournot–Nash equilibrium. *Quarterly Journal of Economics* 98: 185–199.

Stigler, G.J. 1950. Monopoly and oligopoly by merger. *American Economic Review, Papers and Proceedings* 40: 23–34.

Merit Goods

Richard A. Musgrave

Abstract

The term 'merit goods' has no generally agreed application. It is best applied where individual choice is restrained by community values. It may apply also where charity or political redistribution imposes the donors' preferences on recipients; in primary redistribution, society may define fair shares in cash or kind, the latter chosen with regard to what are considered meritorious items for the recipient. However, the concept of merit goods remains within the realm of consumer sovereignty when individuals' 'higher' preferences are imposed on their 'lower' ones.

Keywords

Aristotle; Asymmetric information; Charity; Commitment; Community preferences; Consumer sovereignty; Fair shares; Fiscal theory; Harsanyi, J. C.; Individual choice; Kant, I.; Majority rule; Merit goods; Musgrave, R. A.; Myopia; Paternalism; Pigou, A. C.; Preferences; Primary goods; Private vs. public goods; Subjective vs. ethical preferences; Rawls, J.; Redistribution; Smith, A.; Vickrey, W. S.

JEL Classifications

H4

The concept of merit goods, since its introduction thirty years ago (Musgrave 1957, 1959), has been widely discussed and given divergent interpretations (for surveys, see Head 1966; Andel 1984). Since no patent attaches to the term, it is thus difficult to provide a unique definition. However, most interpretations relate to situations where evaluation of a good (its merit or demerit) derives not simply from the norm of consumer sovereignty but involves an alternative norm. In the following, various situations and their bearing on the concept will be considered.

Merit Goods, Private Goods and Public Goods

While the concept of merit goods was raised in the context of fiscal theory, the term has broader application and should not be confused with that of public (Musgrave 1957, 1959) goods. The distinction between private and public or social goods arises from the mode in which benefits become available, i.e., rival in the one and non-rival in the other case (see ▶ Public Goods). As a result, conditions of Pareto optimality differ, as do the appropriate mechanisms of choice. But whether met through a market or political process, both choices and the normative evaluation of outcomes squarely rest on the premise of individual preference. Consumer sovereignty is taken to apply to both cases. The concept of merit (or, for that matter, of demerit) goods questions that premise. It thus cuts across the traditional distinction between private and public goods. A more fundamental set of issues is raised, issues which do not readily fit into the conventional framework of micro theory as based on a clearly designed concept of free consumer choice.

Pathological Cases

Next, we consider various settings where the norm of consumer sovereignty remains the preferred solution, but where difficulties in implementation have to be met. The most extreme case arises with regard to the mentally deficient or children. In both cases, some guidance is needed and custodial choices have to be made. These, however, may be viewed as exceptional circumstances and not part of the essential merit good problem. It is also evident that rational choice requires correct information, and that the quality of choice is impeded where information is imperfect or misleading. Situations may arise, as in the design of educational programmes, where the quality of choice as eventually valued by the beneficiary's own preference is improved by initial delegation of choice to others whose prior information is superior. Once more,

the implementation of individual preferences is affected, but without questioning their dominance at the normative level.

Other instances arise where rational choice is impeded by oversight or myopia. Individuals, though informed and generally competent to choose, may be inclined to depart from rational choice on certain issues. Thus future consumption tends to be undervalued relative to present consumption (Pigou 1928), while public services may be overvalued because they seem free or undervalued due to dislike of taxation. Rational choice may be impeded in the context of risk-taking, and so forth. Certain goods may thus come to be under or oversupplied for such reasons of misjudgment and their promotion or restriction may be called for. Such situations again pose some departure from the premise of rational choice, but they deal with defects in the implementation of consumer sovereignty, rather than its rejection as a norm.

Rule of Fashion

By assuming individuals to have a well-defined preference structure which may then be interfered with, it is tempting to bypass the fact that individual preferences are not fixed in isolation but are affected by the societal setting in which individuals operate. Taking an extreme view of this dependence (Galbraith 1958), the existence of independent preferences may be denied. Individual preferences become mirror images of fashions in what society approves or holds desirable. But this is too extreme a position. While societal influences enter, they are nevertheless met by individual responses, leaving effective preferences to differ across individuals. Though the preferences of individuals are conditioned by their social environment, own-preferences enter in shaping the individual's responses thereto. It thus seems inappropriate to equate the concept of merit goods with that of fashion.

Community Preferences

As distinct from the rule of fashion, consider a setting where individuals, as members of the community, accept certain community values or preferences, even though their personal preferences might differ. Concern for maintenance of historical sites, respect for national holidays, regard for environment or for learning and the arts are cases in point. Such acceptance in turn may affect one's choice of private goods or lead to budgetary support of public goods even though own preferences speak otherwise. By the same token, society may come to reject or penalize certain activities or products which are regarded as demerit goods. Restriction of drug use or of prostitution as offences to human dignity (quite apart from potentially costly externalities) may be seen to fit this pattern. Community values are thus taken to give rise to merit or demerit goods. The hard-bitten reader regards this as merely another instance of fashion which may be disposed of accordingly. But such is not the case. Without resorting to the notion of an 'organic community', common values may be taken to reflect the outcome of a historical process of interaction among individuals, leading to the formation of common values or preferences which are transmitted thereafter (Colm 1965). As this author sees it, this is the setting in which the concept of merit or demerit goods is most clearly appropriate, and where consumer sovereignty is replaced by an alternative norm.

Paternalism in Distribution

In viewing the problem of individual choice and preferences, we so far have assumed that the individual's endowment from which to choose is given. It remains to consider a set of problems which arise in the context of distribution.

We begin with the case of voluntary giving (Hochman and Rogers 1969). Donor D may derive utility from giving to recipient R, but more so if the grant is specified in kind (e.g., milk) than given in cash (and used for beer). Such paternalistic giving interferes with R's preferences. While R cannot be damaged (the grant can be refused) his or her gain is less than it would be from a cash grant. Charity by way of paternalistic giving thus involves imposition of D's

preferences, of what goods *he* considers of merit for R. At the same time, giving in kind is in line with consumer sovereignty at the donor level, as D's satisfaction depends on what R consumes. Moreover, R cannot suffer a loss, since the grant may be rejected.

A similar problem arises in the context of redistribution through the political process of majority rule. Here, taking as well as giving is involved. While the Rs would prefer to take cash, they may do better by setting for in-kind programmes which appeal to the Ds. Redistribution by majority vote may thus take in-kind form. Once more the Ds may impose their preferences on the Rs, but subject to the terms of the social contract which now permits such intervention via majority rule. Many budget programmes rendering services to the poor (such as health, welfare, and low-cost housing) are of this type, and have indeed come to be classified as merit goods (OECD 1985).

Having considered merit goods in relation to redistribution, it remains to note their bearing on the more basic issue of *primary* distribution. Models of distributive justice have taken a variety of forms, including entitlement to earnings in the Lockean tradition, utilitarian criteria, and entitlement to 'fair shares' (Vickrey 1960; Harsanyi 1955; Rawls 1971). The latter may be viewed in terms of fair shares in income and wealth, while leaving its use to individual choice; or, it may be viewed in terms of a fair share in particular goods or bundles thereof. The role of merit goods arises in the latter context, and indeed bears some relation to the philosopher's concept of 'primary goods'. Moreover, both approaches may be combined in various ways. Thus, society may view it as fair to modify the distribution of income via a tax-transfer scheme, while also arranging the distribution of certain goods (e.g., scarce medical treatment) outside the market rule (Tobin 1970), or society may wish to assure an adequate minimum provision, but do so by providing for a bundle of necessities rather than an equivalent minimum income to be spent at the recipient's choice. Goods separated out for non-market distribution might then be viewed as merit goods.

Multiple Preferences or 'Higher Values'

The reader will note that up to this point we have dealt with settings which, in one way or another, involve some form of departure from the rule of consumer sovereignty. It remains to consider a further perspective, which views the problem within the sovereignty context. This approach postulates that preferences may derive from conflicting sets. This has been noted over the ages, from Artstotle's concept of '*atrasia*,' over the Kantian imperative and Faust's 'two souls' to Adam Smith's impartial observer (Smith 1749). Later the same thought appears in Harsanyi's distinction between subjective and ethical preferences (Harsanyi 1955). A recent illustration follows in Rawls's concept of disinterested choice (Rawls 1971) and Sen's usage of commitment (Sen 1977). The term merit goods has then been applied to goods chosen under the latter ('ethically superior') set of preferences. Such choice may involve private as well as public goods, although they may be more likely to enter in the latter context where they may prove less costly due to the sharing of tax burdens (Brennan and Lomasky 1983).

Conclusion

As the preceding discussion shows, the term merit goods has been applied to a variety of situations. In section "Merit Goods, Private Goods and Public Goods" we have noted that the merit good concept should not be confused with that of public goods. In section "Pathological Cases" we noted that a variety of situations may arise where interference with individual choice is needed but without questioning its validity as the basic norm. In section "Rule of Fashion" we have granted that individual preferences are influenced by social environment, but not to the point of excluding individual-preference based responses. None of these cases offered an appropriate setting in which to apply the merit or demerit concept. The case considered in section "Community Preferences," offering community values as a restraint on individual choice, did, however, fit the pattern and, as I see it, goes to the heart of the merit

concept. Section "Paternalism in Distribution" posed related issues in the context of distribution. Voluntary giving was shown to permit the donor to impose his or her preferences on the donee, and this remains the case, if with lesser force, for political redistribution. Redistribution will tend to be in goods which the donor consider meritorious for the donee. Turning to primary distribution, we noted that society may define fair shares in cash or kind, the latter chosen with regard to what are considered meritorious items for the recipient. Only in section "Multiple Preferences or 'Higher Values'" did use of the merit goods concept remain within the context of the sovereignty norm, dealing now with preferences (merit or demerit wants) of a higher or lower kind. In all, it seems difficult to assign a unique meaning to the term. This writer's preference, as noted before, would reserve its use for the setting dealt with under section "Community Preferences," but that of sections "Paternalism in Distribution" and "Multiple Preferences or 'Higher Values'" may also have a claim.

See Also

- ▶ Public Finance
- ▶ Public Goods

Bibliography

Andel, N. 1984. Zum Konzept der meritorischen Güter. *Finanz Archiv*, New Series 42(3), where extensive literature references are given.

Brennan, G., and L. Lomasky. 1983. Institutional aspects of merit goods analysis. *Finanz Archiv*, New Series 41: 183–206.

Colm, G. 1965. National goals analysis and marginal utility economics. *Finanz Archiv*, New Series 24: 209–224.

Galbraith, K. 1958. *The affluent society*. Boston: Houghton Mifflin.

Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Head, J.C. 1966. On merit goods. *Finanz Archiv*, New Series 25(1): 1–29.

Hochman, H.H., and J.D. Rogers. 1969. Pareto-optimal redistribution. *American Economic Review* 59: 542–557.

Musgrave, R.A. 1957. A multiple theory of budget determination. *Finanz Archiv*, New Series 17: 333–343.

Musgrave, R.A. 1959. *The theory of public finance*. New York: McGraw-Hill.

OECD. 1985. *The role of the public sector*. Paris: OECD.

Pigou, A.C. 1928. *A study in public finance*. London: Macmillan.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Sen, A. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6: 317–344.

Smith, A. 1749. *The theory of moral sentiments*. Reprinted, New York: Liberty, 1969.

Tobin, J. 1970. On limiting the domain of inequality. *Journal of Law and Economics* 13: 263–277.

Veblen, T. 1899. *The theory of the leisure class*. New York: New American Library.

Vickrey, W. 1960. Utility, strategy, and social decision rules. *Quarterly Journal of Economics* 74: 507–535.

Merivale, Herman (1806–1874)

Donald Winch

After a brilliant scholarly career at Harrow, Merivale went on to Oriel College, Oxford, where he obtained a first in classical honours and was elected Fellow of Balliol College in 1828. He was called to the Bar in 1832. In 1837, in succession to Senior, Whately and W.F. Lloyd, he was elected for five years to the Drummond Chair of Political Economy at Oxford. In his introductory lecture he defended political economy from its critics by enunciating the distinction between the science and art of political economy, and by denying that it was based on a degrading view of human nature. But his most permanent contribution to classical political economy was made during the final three years of his tenure of the Chair, when he gave a series of lectures on colonies and colonization which were published in 1841. The success of these lectures opened up for him a new career as a public servant. He became Assistant Under-Secretary for the colonies in 1847, rising to Permanent Under-Secretary in succession to Sir James Stephen in 1848, and later Permanent Under-Secretary to India in 1859. In these

respects he was typical of the new generation of public servants who came to the fore during this period as a result of Britain's growing responsibilities – and acceptance of those responsibilities – in relation to overseas possessions. In addition to these writings on political economy and colonial policy, Merivale wrote on a variety of historical and literary topics for the *Edinburgh Review*, the *Foreign Quarterly*, the *Quarterly Review*, and the *Pall Mall Gazette*. A volume of *Historical Studies* appeared in 1865.

As a political economist, Merivale made no contributions to theory in the fundamental sense, but he did make a number of acute contributions to the disputes surrounding the applications of established economic reasoning to colonies and colonization. Thus in relation to Wakefield's challenge to orthodox Ricardian diagnoses of the British economic situation, Merivale took a middle position, upholding the correctness of Ricardo's theory of declining profit as a long run model, but accepting Wakefield's views as applicable 'under the actual circumstances of society'. In this respect he anticipated the position adopted by John Stuart Mill in his *Principles of Political Economy* (1848). He also pointed out the differences between colonies most likely to benefit and those for whom the principle of inhibiting access to public land along Wakefieldian lines would be inappropriate. The distinction turned on the availability of export markets and the consequent need for regular labour supplies: in other words, on whether colonial agriculture was predominantly market-oriented. Where this was not the case, as in the northern and mid-western states of America at that time, the Wakefield principle was likely to encumber the opening up of fertile territories without yielding comparable economic advantages to society at large. The lectures, together with the reflections added in 1861, after Merivale had acquired experience as a senior civil servant, contain a balanced commentary on the whole range of colonial policy in this period, including the likely incompatibility of self-government in the colonies with the continuance of imperial control, slavery and the treatment of native peoples, and Britain's imperial 'mission'.

See Also

▶ Colonies

Selected Works

1861. Lectures on colonisation and colonies delivered before the University of Oxford in 1839, 1840 and 1841. Oxford. Reprinted, 1928.

Bibliography

Ghosh, R.N. 1967. *Classical macroeconomics and the case for colonies*. Calcutta: New Age Publishers.
Winch, D. 1965. *Classical political economy and colonies*. London: Bell and Sons.

Merton, Robert C. (Born 1944)

Darrell Duffie

Abstract

Robert C. Merton, who developed the theory of option pricing with Myron Scholes and the Fischer Black, is responsible for a new approach to investments and asset pricing, based in part on stochastic calculus. Awarded the Nobel prize in 1997, Merton's other contributions to financial economics include the intertemporal capital asset pricing model (ICAPM). He has written extensively on pension planning, social security, and bank deposit insurance.

Keywords

American Finance Association; Arbitrage; Black, F.; Black–Scholes option pricing model; Derivative securities; Intertemporal capital asset pricing model; Long-Term Capital Management (LTCM); Markowitz, H.; Mean-variance investment theory; Merton, R.; Miller, M.; Modigliani, F.; Rational option pricing theory; Scholes, M.; Sharpe, W.; Stochastic calculus

JEL Classifications
B31

Robert C. Merton, awarded the 1997 Nobel Memorial Prize in Economics, was born in New York City on 31 July 1944. His father, Robert K. Merton, was a noted sociologist, to say the least. This biographical sketch of Robert C. Merton and his contributions to financial economics may seem brief, given the gigantic impact that he had on economics and financial-market practice.

Merton's university education veered from applied mathematics at Columbia University (BS, 1966) and the California Institute of Technology (MS, 1967) to economics at MIT (Ph.D., 1970), where he quickly joined Paul Samuelson as student, then research assistant, faculty colleague, and collaborator. Their paper on warrant pricing (1969a) hinted at Merton's later massive contributions to 'the optionpricing formula' and to dynamic investment theory, which followed almost immediately. Within a few years of his arrival at MIT in 1967, it is no exaggeration to say that Merton had transformed his newly chosen field of financial economics and, more broadly, dynamic modelling in economics.

Only a decade before Merton framed his revolutionary new approach to financial modelling, Modigliani and Miller (1958) had used arbitrage reasoning to discover the irrelevance of corporate capital structure and dividend policy in perfect capital markets. About five years before Merton came on to the scene, William Sharpe (1964) had adapted Markowitz's mean-variance investment theory to establish the relationship between risk and expected return in market equilibrium. These pre-Merton breakthroughs were based on static reasoning. Merton exploited stochastic calculus – a completely new approach to dynamic modelling under uncertainty – in order to extend these insights and to open entirely new paths of discovery. The crucial tool of stochastic calculus that Merton brought into financial modelling is the formula of Kiyoshi Itô (1951), whereby, under suitable technical regularity, the rate of change of the conditional expectation of $f(X(t), t)$, for an Itô process X and a smooth function $f(\cdot\cdot)$, is

$$f_t(X(t),t) + f_x(X(t),t)m(t) + \frac{1}{2}f_{xx}(X(t),t)v(t), \quad (1)$$

where $m(t)$ is the rate of change of the conditional expectation of $X(t)$, and $v(t)$ is the rate of change of the conditional variance of $X(t)$. (Subscripts indicate partial derivatives.)

Consider, for example, Merton's approach (1969b; 1971) to investment, in which $X(t)$ is the wealth of a risk-averse investor whose current optimal conditional expected utility for final wealth is $f(X(t), t)$. (The conjectured dependence of indirect utility on wealth and time only is tantamount to the independence over time of asset returns, which Merton relaxed in 1973b.) The current portfolio p of investments determines the 'local mean' $m(t, p)$ and 'local variance' $v(t, p)$ of changes in wealth. At anything other than an optimal portfolio strategy, Bellman's principle of optimality implies that the conditional expected change of $f(X(t), t)$ is negative, so (1) suggests that

$$\mathrm{Max}_p f_t(X(t),t) + f_x(X(t),t) \\ m(t,p) + \frac{1}{2}f_{xx}(X(t),t)v(t,p) = 0. \quad (2)$$

Because the mean $m(t, p)$ and variance $v(t, p)$ of the 'local return' on wealth are linear and quadratic, respectively, with respect to the portfolio choice p, the firstorder optimality conditions for (2) provide an explicit solution for p in terms of the derivatives of $f(\cdot\cdot)$. Substitution of this solution for p into the same equation (2) leaves a partial differential equation to solve for $f(x, t)$. Merton was able to give explicit solutions in certain cases. For example, with expected power utility for final wealth, the indirect utility must inherit the same degree of homogeneity with respect to wealth. Merton's problem is still the classic textbook example of stochastic control to which graduate students in finance and other fields, even beyond economics, are first exposed. The associated insights into lifetime investment planning are striking, and have led to an immense literature of extensions.

Although this is no place to derive it, the Black–Scholes formula $f(x, t)$ for the price at time t of an option on an asset whose current

market value is x is similarly obtained by the 'risk-neutral' valuation equation

$$f_t(x,t) + f_x(x,t)rx + \frac{1}{2}f_{xx}(x,t)\sigma^2x^2 = rf(x,t), \quad (3)$$

where r is the continuously compounding risk-free borrowing rate and σ is the volatility (the standard deviation of annualized continuously compounding returns) of the underlying asset. The boundary condition for (3), in the case of a call option with an exercise date T and exercise price K, is $f(x, T) = max(x - K, 0)$, because this is the market value of the right, but not the obligation, to buy the stock for K when it trades in the market for x. Black and Scholes (1973) solved this equation with the famous formula named for them. For the market value of a general contingent claim paying $g(X(T))$ at T, the same differential equation (3) applies under technical conditions, with the boundary condition $f(x, T) = g(x)$.

By virtue of Itô's formula (1), one can view (3) as a statement that the option's expected rate of return may be treated as the risk-free rate of return, provided that we replace the actual mean rate of return on the underlying asset with the risk-free rate. Indeed, this is roughly how Black and Scholes (1973) interpreted their original derivation of (3), which was based on a particular general-equilibrium model. Merton, however, noted that changes in the market value of the option over time could actually be replicated by trading the underlying asset, financing any cash needs with risk-free borrowing. This 'arbitrage' strategy leads to (3) without reference to a particular general-equilibrium model, since arbitrage is ruled out in any equilibrium. Black and Scholes acknowledged Merton for this alternative approach, which was the genesis of both an enormous academic literature on contingent claims pricing and the professional practice of 'financial engineering', a field that includes a vast array of financial pricing and risk-management methods.

Among the most influential applications that Merton developed on the basis of his approach to derivative asset pricing was his insight (1974) that the equity and debt of a corporation may be viewed as derivative securities written on the assets of the firm, and priced accordingly. This idea was developed independently in Black and Scholes (1973). In any case, this widely known 'Merton model of corporate debt' is the basis of much modern fundamental market analysis of corporate debt and credit derivatives, in practice and academic research, including both pricing and default prediction.

Rounding out the series of major results that Merton produced within a stunningly short period of time were his intertemporal capital asset pricing model (ICAPM) (1973b) and his theory of rational option pricing (1973a). Merton's ICAPM extended Sharpe's CAPM to a dynamic framework, relieving it of its dependence on mean-variance utility because, from (2), we see that only the ('instantaneous') mean and variance of returns matter for conditional mean rates of change of utility, under technical conditions. More importantly, the ICAPM showed how the expected returns of assets in a multi-period setting compensate not only for exposure to the risk associated with the return of the market portfolio but also for exposure to the risks associated with changes in state variables determining future conditional distributions of asset returns. These latter risks introduce hedging motives not present in a static model. Merton's Theory of Rational Option Pricing (1973a) shored up the foundations of the Black–Scholes option pricing model and treated a variety of related issues, in particular a rational approach to exercising and pricing American options. A few years later Merton (1977) provided deeper foundations for the basic arbitrage reasoning underlying the pricing of derivatives by replacing his earlier 'instantaneous return' arbitrage argument, the original basis of the Black–Scholes formula, with the construction of dynamic portfolio trading strategies that specified, at each state and date, the actual quantities of each type of security that an investor would hold in order to replicate the final payoff of the target contingent claim.

After 1978 Merton shifted his attention from foundational theories of investment and asset pricing to applications of those theories, paying

special attention to the institutional features of financial markets and to related issues of public policy. For example, a series of papers addressed pension planning, social security, and bank deposit insurance. He also worked on corporate capital budgeting, labour contracts, financial intermediation, and the risk management of financial institutions, among many other applications. Merton even turned his hand to some empirical research on investments. His 1987 presidential address to the American Finance Association raised some influential new ideas regarding the impact of market imperfections and incomplete information on equilibrium asset prices.

In 1988 Merton moved from MIT, of whose faculty he had been a member since 1970, to Harvard University. While he has maintained direct involvement in financial markets in various capacities throughout his professional career, for example as a consultant, in 1993 Merton took a more significant step in this direction by becoming one of the first principals of the now notorious hedge fund, Long-Term Capital Management (LTCM). In its first years, the great financial successes of LTCM were attributed in large measure to the unusually deep team of talented financial minds, notably including both Merton and Myron Scholes, which had been assembled by John Merriwether, LTCM's founder. When LTCM failed spectacularly in 1998, some pundits ironically blamed undue reliance on sophisticated financial modelling, in some cases singling out Merton and Scholes. The record, however, seems to point to initial successes based on high leverage, attractive financing, and good trading, and then failure caused by high leverage coupled with the results of some unwise or unlucky trading, exacerbated by a 'rush to the exits' by other investors who held large positions similar to those of LTCM. In 2002, Merton co-founded Integrated Finance, a financial advisory firm.

As of this writing, Merton continues to publish and speak influentially, and remains on Harvard's faculty. In addition to the Nobel Prize, Merton is the recipient of numerous awards and honorary degrees, and is widely viewed as one of the alltime most respected leaders and researchers of his profession.

See Also

- ▶ Contingent Valuation
- ▶ Miller, Merton (1923–2000)
- ▶ Modigliani, Franco (1918–2003)
- ▶ Scholes, Myron (Born 1941)
- ▶ Sharpe, William F. (Born 1934)
- ▶ Social Insurance
- ▶ Social Security in the United States

Selected Works

1969a. (With P. Samuelson.) A complete model of warrant pricing that maximizes utility. *Industrial Management Review* 10(Winter): 17–46. Chapter 7 in *Continuous-time finance*.

1969b. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics* 51: 247–257. Chapter 4 in *Continuous-time finance*. 1992.

1971. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3: 373–413. Chapter 5 in *Continuous-time finance*, 1992.

1973a. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183. Chapter 8 in *Continuous-time finance*, 1992.

1973b. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887. Chapter 15 in *Continuous-time finance*, 1992.

1974. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470. Chapter 12 in *Continuous-time finance*, 1992.

1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3: 125–144. Chapter 9 in *Continuous time finance*, 1992.

1977. On the pricing of contingent claims and the Modigliani–Miller theorem. *Journal of Financial Economics* 5: 241–249. Chapter 13 in *Continuous time finance*, 1992.

1978. On the cost of deposit insurance when there are surveillance costs. *Journal of Business* 51: 439–452. Chapter 20 in *Continuous-time finance*, 1992.

1980. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics* 8: 323–361.
1983a. On the role of social security as a means for efficient risk-bearing in an economy where human capital is not tradeable. In *Financial aspects of the US pension system*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press.
1983b. On consumption-indexed public pension plans. In *Financial aspects of the US pension system*, ed. Z. Bodie and J. Shoven. Chicago: University of Chicago Press 1983. Chapter 18 in *Continuous-time finance*, 1992.
1985. Implicit labor contracts viewed as options: A discussion of 'Insurance aspects of pensions'. In *Pensions, labor, and individual choice*, ed. D. Wise. Chicago: University of Chicago Press.
1987a. (With Z. Bodie and A. Marcus.) Pension plan integration as insurance against social security risk. In *Issues in pension economics*, ed. Z. Bodie, J. Shoven and D. Wise. Chicago: University of Chicago Press.
1987b. A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42: 483–510.
1987c. (With Z. Bodie and A. Marcus.) Defined benefit versus defined contribution pension plans: What are the real tradeoffs? In *Pensions in the U.S. economy*, ed. J. Shoven and D. Wise. Chicago: University of Chicago Press.
1990. The financial system and economic performance. *Journal of Financial Services Research* 4: 263–300.
1992. *Continuous-time finance*, Rev. edn. Cambridge, MA: Basil Blackwell.
1993a. (With Z. Bodie.) Deposit insurance reform: A functional approach. *Carnegie–Rochester Conference Series on Public Policy* 38: 1–34.
1993b. (With A. Perold.) Theory of risk capital in financial firms. *Journal of Applied Corporate Finance* (Fall): 16–32.
1993c. (With Z. Bodie.) Pension benefit guarantees in the United States: A functional analysis. In *The future of pensions in the United States*, ed. R. Schmitt. Philadelphia: University of Pennsylvania Press.
1995. Financial innovation and the management and regulation of financial institutions. *Journal of Banking and Finance* 19: 461–481.
1997. A model of contract guarantees for credit-sensitive, opaque financial intermediaries. *European Finance Review* 1: 1–13.
2000. (With Z. Bodie.) *Finance*. New Jersey: Prentice-Hall.
2005. (With Z. Bodie.) The design of financial systems: Towards a synthesis of function and structure. *Journal of Investment Management* 3: 1–23.

Bibliography

Bernstein, P. 1992. *Capital ideas*. New York: Free Press.
Black, F. 1989. How we came up with the option formula. *Journal of Portfolio Management* 15: 4–8.
Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
Duffie, D. 1998. Black, Merton and Scholes – Their central contributions to economics. *Scandinavian Journal of Economics* 100: 411–424.
Greenspan, A. 1998. Private-sector refinancing of the large hedge fund, Long-Term Capital Management. Testimony before the Committee on Banking and Financial Services, US House of Representatives, 1 October. Online. Available at https://www.federalreserve.gov/boarddocs/testimony/1998/19981001.htm Accessed 26 July 2005.
Itô, K. 1951. On a formula concerning stochastic differentials. *Nagoya Mathematics Journal* 3: 55–65.
Jarrow, R., and P. Protter. 2004. A short history of stochastic integration and mathematical finance: The early years, 1880–1970. In *The Herman Rubin festschrift,* IMS Lecture Notes 45, 75–91. Institute of Mathematical Statistics: Bethesda, MD.
Markowitz, H. 1959. *Portfolio selection: Efficient diversification of investment*. New York: Wiley.
Modigliani, F., and M. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48: 261–297.
Royal Swedish Academy of Sciences. 1998. The nobel memorial prize in economics 1997. *Scandinavian Journal of Economics* 100: 405–409.
Schaefer, S. 1998. Robert Merton, Myron Scholes and the development of derivative pricing. *Scandinavian Journal of Economics* 100: 425–445.
Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Methodenstreit

Daniel R. Fusfeld

Keywords

Hobson, J. A.; Institutional economics; Menger, C.; Methodenstreit; Methodology of economics; Schmoller, G. von; Veblen, T.

JEL Classifications

B1

The 'battle of methods' between Carl Menger (1840–1921) and Gustav Schmoller (1838–1917) is one of the most important methodological debates in the history of economics. It began with the publication of Menger's book on method (1883), which made the case for pure theory based on assumptions about behaviour and antecedent conditions. Schmoller responded with a strongly worded review (1883) that argued for principles of economics based on empirical historical data and the inductive method. Menger answered with an equally vehement statement of *The Errors of Historicism* (1884). The infuriated Schmoller refused even to read it (Schmoller 1884). A torrent of books and papers by others followed over the next several decades. The best summary of the entire controversy is Ritzel (1951).

Like most disputes over method in economics, the opposing views were related to more complex disagreements over the nature and scope of economics and its policy implications. Menger's assumptions about behaviour implied a social system composed of selfishly motivated individuals; Schmoller assumed the existence of individuals grouped into nations, with group as well as individual goals. More important, Menger's conclusions emphasized the primacy of laissez-faire policies designed to allow as large a scope as possible to market adjustment processes. Schmoller's conclusions supported the interventionist and state-building policies of the newly unified German nation. In addition, the Ministry of Education in Berlin gave almost exclusive preference to the Schmoller school in appointing university professors. Menger was attacking the 'official' economics prevailing in Germany and its almost monopolistic control over university appointments. In addition to economic method, academic freedom and the role of the state were at issue.

On the basic issue of the place of theory and empirical studies in economics, Menger and Schmoller agreed that both were necessary. They disagreed, however, on the emphasis to be placed on each and their role in the development of conclusions. Menger argued that 'pure' economics based on assumptions of wide and perhaps universal generality, could be developed through correct logical analysis to arrive at conclusions of equally broad applicability and usefulness. Propositions based on empirical data, however, would be correct only for the limited data on which they were based. Since empirical data were always partial, as well as bounded by time and space, the conclusions drawn from them must be both problematic and of limited generality. Correct and general propositions could be derived through rigorous logic from assumptions not bounded by time, space or special circumstance, however.

Empirical studies entered Menger's method in two ways. First, they could be used to verify or illustrate the results of theoretic inquiry. Second, they were necessary when theoretic principles were applied to specific instances or policy problems. Empirical studies were required to define the situation to which theoretic principles were applied, and to delimit the applicability of the conclusions. Data acted as a bridge between the principles of pure economics and the policy problems of applied economics. Indeed, Menger warned against application of pure theory to applied problems without thorough empirical studies.

Schmoller also advocated use of both empirical studies and theory, but in a different combination. He rejected Menger's logical deductive method for three chief reasons: its assumptions were unrealistic, its high degree of abstraction

made it largely irrelevant to the real-world economy, and it was devoid of empirical content. The theory was therefore useless in studying the chief questions of importance to economists: how have the economic institutions of the modern world developed to their present state, and what are the laws and regularities that govern them? The proper method was induction of general principles from historical–empirical studies (Schmoller 1883). In the Hegelian tradition of 19th-century German scholarship, Schmoller conceived of the economy as a dynamic and evolving set of interrelated institutions whose laws of development could not be understood in terms of an abstract theory of constrained choice. One reason for the polarized arguments of the Methodenstreit was that the disputants were talking about different things.

How were the historical laws of economic development and change to be determined? Schmoller was not clear on that point, although he devoted five chapters of the introductory section of his *Grundriss* to a survey of the history and method of economics (Schmoller 1900–4). The starting point of his method was empirical research rather than assumptions. The second step was to organize the data in a logical fashion, to bring out the essential nature of economic phenomena. The third step was to identify the relationship between phenomena in the context of their continually changing interaction and development. At all stages of the inquiry, empirical research was to be used to obtain the propositions of steps two and three. The connecting link between data and generalizations was not spelled out, although in retrospect we can interpret the procedure as an early version of the gestalt method and the use of pattern models.

The Methodenstreit had a significant impact on the development of economics. Schmoller's attack on the logical deductive method as inherently devoid of empirical content coincided with similar critiques by the British economic historians, John A. Hobson and the American institutionalists led by Thorstein Veblen. These critics forced the adherents of neoclassical economics to bring empirical studies more fully into the mainstream of economic thought and practice. After the Methodenstreit a combination of theory and empirical studies was almost universally accepted by economists as necessary.

Menger's method of combining them was adopted, however. In the 20th century economics became increasingly a theoretic discipline based on 'as if' assumptions, which are developed by rigorous logical methods to derive general propositions.

Hypotheses about reality, derived from the general propositions, are then tested against empirical studies. Schmoller's vision of an empirical discipline based on factual studies, in which generalizations are both derived from and tested against data as they are developed, remains only among critics of the mainstream in a new battle of methods that has erupted a hundred years later.

See Also

- ▶ Historical School, German
- ▶ Institutional Economics
- ▶ Menger, Carl (1840–1921)
- ▶ Methodology of Economics
- ▶ Schmoller, Gustav von (1838–1917)

Bibliography

Menger, K. 1883. *Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Oekonomie insbesondere*. Berlin: Duncker & Humblot. Ed. L. Schneider and trans. F.J. Nock as *Problems of economics and sociology*, Urbana: University of Illinois Press, 1963.

Menger, K. 1884. *Die Irrthümer des Historismus in der deutschen Nationalökonomie*. Vienna: Alfred Hölder.

Ritzel, G. 1951. *Schmoller versus Menger*. Offenbach: Bollwerk-Verlag.

von Schmoller, G. 1883. Zur Methodologie der Staats- und Sozialwissenschaften. *Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft im deutschen Reich* 8: 974–994.

von Schmoller, G. 1884. *Jahrbuch für Gesetzgebung, Verwaltung und Volkwirtschaft in deutschen Reich*, 677.

von Schmoller, G. 1900–4. *Grundriss der allgemeinen Volkswirtschaftslehre*. Leipzig: Duncker & Humblot.

Methodological Individualism

Kaushik Basu

Abstract

Methodological individualism holds that a proper explanation of a social regularity or phenomenon is grounded in individual motivations and behaviour. Although many economists claim to be methodological individualists, economics has always used social concepts and categories. As Schumpeter pointed out, nearly a century ago, price in a competitive model is an irreducibly social concept. Each individual takes the price as given but the price that comes to prevail is an outcome of the choices made by all individuals. Since Veblen, economists have increasingly recognized that individual preferences are endogenous and may be responsive to what happens in society at large.

Keywords

Arrow, K.; Atomism; Bertrand oligopoly; Buchanan, J.; Class interest; Cournot oligopoly; Dworkin, R.; Endogenous preferences; German school; Hayek, F. A.; Homo economicus vs. homo sociologicus; Menger, C.; Methodological holism; Methodological individualism; Normative individualism; Schumpeter, J.; Self-interest; Smith, A.; Veblen, T.; Weber, M.

JEL Classifications

B4

Methodological individualism is a doctrine in the social sciences according to which a proper explanation of a social regularity or phenomenon is one that is grounded in individual motivations and behaviour. In other words, according to this methodology, individual human beings are the basic units from which we must build *up* in order to understand the functioning of society, economy and polity. We may not in all our research succeed in doing so but to committed methodological individualists such research must be viewed as provisional and ideally be accompanied by a slight feeling of inadequacy on the part of the researcher.

The social scientists who have been the focus of much of the debate on methodological individualism and, paradoxically, also the ones least touched by the debate are economists. Economists are typically held up as examples of the most unbending methodological individualists; and, on the rare occasions when economists have joined this debate, they have tended to agree with this. The difference is that most non-economists mean this as criticism, whereas most economists take it as praise.

At first sight this characterization of economics seems right. Textbooks of microeconomics almost invariably begin by specifying individual utility functions or preference relations and asserting that human beings are rational in the sense that they behave so as to maximize their own utilities. They then build up from this to explain market phenomena, make claims about social welfare and discuss prospects of national economic growth. In some macroeconomic models economists are unable to build all the way up from individual behaviour and use aggregate behaviour descriptions as the starting point. But these models are almost always accompanied by an effort to 'complete' them with proper micro-foundations; and the profession regards these models as somewhat incomplete and awaiting the definitive work.

That economics may not actually be quite as methodologically individualistic as often presumed by both the discipline's admirers and its critics is a matter to which I return below. What is interesting to note here is that the debate on methodological individualism has been a surprisingly cantankerous one that has spawned enemies and intrigues. Some social scientists have sworn by it: no other method is worth its salt. Others have castigated it as an instrument of exploitation and maintenance of the status quo. Concepts and categorizations have multiplied over the years. We have come across methodological holism, methodological solipsism, atomism, 'MIs' (that is,

methodological individualisms) of different types – 1, 2, 3. . . – creating the impression that the British intelligence had somehow got involved in the quest to understand this elusive concept.

One cause of the controversy is the confounding of positive and normative social science. To some commentators, methodological individualism implies that it is fine to leave it all to individuals, and by implication it amounts to an argument against government intervention. Friedrich von Hayek (1942) and James Buchanan (1989), for instance, have taken this line, as have some sociologists, who felt that the conservatism of traditional economics is founded in its adherence to methodological individualism. But this happens because of a possibly logical error, a failure to appreciate Hume's law, namely, that a normative proposition cannot be derived from a purely positive analysis. Kenneth Arrow (1994) has rightly criticized the tendency of some writers to treat methodological individualism and 'normative individualism' as inextricably linked. Similarly, Marxists often link methodological individualism automatically to certain ethical implications. Roemer (1981) and Elster (1982) argue that this is not a valid link. In what follows I treat the two as separate and assume that methodological individualism has no automatic normative implications.

Origins

The term 'methodological individualism' was probably used for the first time in the English language in 1909 by Joseph Schumpeter. Even if that is not so, Schumpeter certainly thought so, and he pointed out in his paper in the *Quarterly Journal of Economics* that year that he had actually coined the term in German the previous year. But methodological individualism had been *practised* from much earlier, at least as early as Adam Smith (1776), and was described as a deliberate methodology, though without the term itself being used, by Carl Menger in 1883 (Menger 1883). Max Weber's later exposition of it was published posthumously in 1922 (Weber 1922).

From the perspective of economics it seems reasonable to treat Menger as the first proponent of methodological individualism. He did so vociferously, dismissing the German historical school of economists and their methods as outdated and flawed. He advanced the idea of 'spontaneous order' in society, which sprang from atomistic individual behaviour, reminiscent of Adam Smith's 'invisible hand' and the efficiency of markets that was an outcome of rational, self-interested behaviour on the part of individuals. Menger not only failed to acknowledge that some of his ideas on spontaneous social order were already there in Adam Smith but wrote in a tone almost suggesting that Smith had taken those ideas from him.

A distinction is often drawn in philosophy between methodological individualism and 'atomism'. The latter is treated as a more extreme version of individualism, in which it is possible to characterize each individual fully without reference to society and then explain social behaviour by simply imagining such individuals being brought together in one society. Since the proponents of these ideas did not really define terms with that much care – and when they did, they went on to write in a way that disregarded their own definitions – I shall refrain from drawing fine distinctions and treat these neighbouring terms as all representing the broad *idea* of methodological individualism. Moreover, concepts like these are probably innately indefinable. They are understood through a combination of approximate definitions and repeated use.

It is useful in an exposition like this to think of the polar opposite of the term under consideration. This is captured in the concept of 'methodological holism', developed (without endorsement) by the philosopher John Watkins. Methodological holism is the belief that there are 'macroscopic laws that are *sui generis* and apply to the system as an organic whole' (Watkins 1952, p. 187), and the behaviour of its components had to be deduced from it. In economics, this would imply beginning our analysis by stating the laws of an aggregate economy and, perhaps, the behaviour of prices and industries and, from that, deducing how individuals behaved and what motivated them. Stated

in these terms, it immediately becomes clear from a perusal of almost any microeconomics textbook that economics belongs essentially to the methodological individualist end of the spectrum defined by methodological holism at one end and individualism at the other.

After these writings, interest in the subject flagged. Social scientists, especially economists, continued to do research without trying to explicitly articulate the method that they were in fact using. The feeling developed among economists that the issue of methodological individualism was either trivial or had been resolved in their favour.

In the early 1990s the economists' gathering insouciance was challenged by Rajeev Bhargava (1993) and Kenneth Arrow (1994). Bhargava summarizes various points of view on the subject and then challenges the orthodoxy, especially within economics. But he also expresses well the philosopher's inevitable anxiety in a debate like this, which stems from not knowing whether what one is grappling with is something profound or trivial. As he writes, 'On reading the literature one is swung between exuberance and despair, from feeling that all problems have been resolved to one that none has ... Gradually an intense frustration overwhelms the reader: perhaps there was nothing worth discussing in the first place. What on earth was all the fuss about?' (1993, p. 5).

What he settles for as the best face of methodological individualism is 'intentionalism'. The intentional man is somewhere between the imaginary *homo economicus* and equally rare *homo sociologicus*. He can choose and decide individually but he is not a relentless, maximizing agent. He has psychology and a sense of social norms, which get in the way of selfish maximization. Bhargava then develops the idea of 'contextualism' as a challenge to methodological individualism, including intentionalism. The challenge consists of arguing that a variety of beliefs and practices in everyday life make sense only in the *context* of the society where they occur. Hence, in describing a society or an economy we are compelled to use concepts which are *irreducibly social*.

The reason why the assertion that certain beliefs and concepts are inextricably social is unlikely to stir a hornet's nest is that, although many economists claim to be rigid adherents of methodological individualism, they do use and have always used social concepts and categories. This is convincingly argued by Arrow. He points out how a variable such as price in a competitive model is an irreducibly social concept. Each individual takes the price to be given but the price that comes to prevail is an outcome of the choices made by all the individuals. So economists constructing equilibrium models, who claim to be hardened methodological individualists, are actually not so, at least in the sense that they use some concepts that are irreducibly social. Knowingly or unknowingly they follow a method which uses social categories. In fact, this was explicitly recognized by Schumpeter in his classic essay on methodological individualism, where he noted 'prices are obviously social phenomena' (1909, p. 217).

Preferences and Groups

There are more contentious claims that one can make about the role of social concepts in economics. One of these relates to the *permissibility* of a certain class of propositions in social science, such as: 'The landlord will undertake action A, *because* it is in the landlord's class interest to do so.' (Action A could, for instance, be: 'refuse to hire a servant who has fled another landlord's employment and offers to work for this landlord for a low wage'). Let me call this proposition P.

The bone of contention between neoclassical and traditional Marxist economists is frequently whether such propositions are permissible. Many neoclassical economists and some political scientists (especially those belonging to the positive political economy school) believe that P is not permissible – a person's class interest must be not be treated as an innate characteristic in the same way that his self-interest may be. A small group of writers maintain that even

Marxism is compatible with methodological individualism and that class and other aggregate behaviour should, ideally, be built from individual motivations and preferences (Roemer 1981; Elster 1982).

In any case, whether or not proposition P is wrong, mainstream economics certainly considers it so. If an economist were to use an axiom like proposition P, she would usually want to first satisfy herself why it may be in the landlord's *self*-interest to behave in a way which is in his *class* interest. However, this does not negate the use of beliefs and other concepts and variables which are irreducibly social. It is not clear whether a researcher who does both (that is, resists explaining individual behaviour solely in terms of its ability to serve group or class interests but uses concepts and beliefs which are inherently social) is a methodological individualist. But this is a purely definitional matter and of no great consequence. The important and contestable question is whether assumptions like proposition P should or should not be used. I take the view that it is best to avoid such assumptions as far as possible, without making that into a dogma.

There are some fundamental ways in which modern economics has moved further away from methodological individualism than merely by using irreducibly social concepts, like prices, and even without using propositions, like P. I here mention two. First, most models of economics make use of the idea of 'rules of the game'. In Cournot oligopoly, firms choose quantities and then wait for prices to form. In Bertrand oligopoly, firms set prices and then wait to sell what the market demands from them. In most real-life situations, these rules evolve over time through intrinsically social processes. We may not fully understand what these social processes are, but few individuals will deny their existence. Arrow (1994) has emphasized this and also the importance of '*social* knowledge.'

Second, there is increasing recognition in economics that individual preferences are endogenous. They evolve over time and may be responsive to what happens in society at large. As Thorstein Veblen (1899) recognized, around the time when neoclassical economics was taking shape, human preferences for certain objects often depend on who else is consuming those objects. If a film star wears a brand-name shirt, you may be willing to pay more for that same shirt. If the elite likes a particular wine, then some people will acquire a taste for that wine; moreover, such people will be viewed by others as belonging to the elite because of their taste in wine. In other words, people often use goods to associate themselves with other people who use those goods (Basu 1989). These are obvious matters (though they were sidelined during the time of Veblen) and any economist whose ability to think is not damaged by excessive textbook education will recognize that these kinds of preference endogeneity exist. What is remarkable about Schumpeter's (1909) essay is that he understood (admittedly in a somewhat inchoate way) that this recognition may cut into the methodological individualism of economics. He observed how, given the human tendency to conform to society, 'there will be a tendency to give [each individual's] utility curves shapes similar to those of other members of the community' (1909, p. 219).

To see how this can ruffle methodological individualism, suppose that each person likes to wear jeans if more than 60 per cent of society wears jeans; more precisely, suppose that, if over 60 per cent of society wears jeans, each person is willing to pay for jeans more than the marginal cost of producing them; otherwise they are willing to pay less. This society will have two possible equilibria: one in which no one wears jeans and another (however revolting it may be to visualize this) in which everyone wears jeans. In models of this kind there is an interdependence between society's behaviour and each individual's preference. Once we recognize this, there is no reason to start our analysis by characterizing the individual. We may still do so through force of habit. But we could equally begin by considering a social behaviour postulate – for instance, that 50 per cent of people wear jeans. Then we work out how much each individual prefers to wear jeans (and so how much each is willing to pay for his or her jeans) and check whether the initial social postulate is sustainable. If it is, then we have found an

equilibrium. If not (as in the above example), then the behaviour is not one that will prevail in equilibrium. This method is one of neither methodological individualism nor methodological holism. It is therefore evident that, as economics becomes more sophisticated, it is moving away from pure individualism towards this kind of a hybrid methodology.

Normative Statements

An interesting and unexpected area where methodological individualism is violated is in some of our normative statements. We often pass moral judgement on groups of people which cannot be reduced to the individuals in the group. Normative propositions of the following kind are common: 'It is a shame that no one in your university does research on poverty.' If you asked the person making this observation whether he was blaming you for not doing research on poverty, he would typically claim that he was not; in fact, he would deny that he was casting aspersions on *any* individual but criticizing the collectivity of individuals in the university. This amounts to an implicit rejection of individualism.

Methodological individualism in the context of normative statements like the above one has not been much analysed, but Ronald Dworkin has provided an interesting analysis. He argues that in situations of group responsibility it may be reasonable to *personify* the group. Thus, when a corporation produces a dangerously defective good but it is not possible to pin down the responsibility on any particular individual, we may need to treat the corporation as a moral agent and apply 'facsimiles of our principles about individual fault and responsibility to it' (Dworkin 1986, p. 170). And *then*, by virtue of the *corporation's* responsibility, we may proceed to hold some or all of the individual *members* of the corporation responsible. This is interesting because it comes as close to Watkins's 'methodological holism' as we are likely to encounter anywhere. Individuals are still essential units in Dworkin's analysis but, unlike in standard methodological individualism, judgement of the group *precedes* judgement of the individual.

Dworkin argues that we unwittingly often use this method. This happens when we talk of the state's responsibility for certain kinds of individual rights. Thus we talk of the state's obligation to ensure that no one is assaulted by others. Moreover, we do this even before agreeing on how this responsibility is to be apportioned across various units and agents of the state, such as the police and the bureaucracy. Dworkin (1986, p. 171) points out how we discuss the community's responsibility and 'leave for *separate* consideration the different issue of which arrangement of official duties would best acquit the communal responsibility' (emphasis added).

It is possible to criticize Dworkin's line (see Basu 2000) by arguing that the personification of the corporation or the community has to be an interim construct. It will be sustained if we *can* then apportion the blame among the members of the corporation. If, however, we find that we are not able to spread the blame among the individuals in some reasonable way, then we may have to forgo our initial stand, which held the corporation responsible, or at least maintain that there is no way to take the next step of tracing the fault to individuals.

Interestingly, this brings us back to the kind of analysis defended in the case of endogenous preferences. And this suggests, once again, that what is needed for modern social science is neither holism not individualism but a hybrid methodology that, at least for now, lacks a name.

See Also

- ▶ Collective Rationality
- ▶ Economic Man
- ▶ Individualism Versus Holism
- ▶ Social Norms

Bibliography

Arrow, K. 1994. Methodological individualism and social knowledge. *American Economic Review* 84: 1–10.

Basu, K. 1989. A theory of association: Social status, prices and markets. *Oxford Economic Papers* 41: 653–671.
Basu, K. 2000. *Prelude to political economy: A study of the social and political foundations of economics*. Oxford: Oxford University Press.
Bhargava, R. 1993. *Individualism in social science: Forms and limits of methodology.* Oxford: Oxford University Press.
Buchanan, J. 1989. The state of economic science. In *The state of economic science*, ed. W. Sichel. Kalamazoo: Upjohn Institute for Employment Research.
Dworkin, R. 1986. *Law's empire*. Cambridge, MA: Harvard University Press.
Elster, J. 1982. Marxism, functionalism and game theory. *Theory and Society* 11: 453–482.
Menger, C. 1883. *Investigations into the method of the social sciences with special reference to economics.* English translation. New York: New York University Press, 1986.
Roemer, J. 1981. *Analytical foundations of Marxian economic theory.* Cambridge: Cambridge University Press.
Schumpeter, J. 1909. On the concept of social value. *Quarterly Journal of Economics* 23: 213–232.
Smith, A. 1776. *Inquiry into the nature and causes of the wealth of nations*, 1981. Indianapolis: Liberty Classics.
Veblen, T. 1899. *The theory of the leisure class*. London: Macmillan.
Von Hayek, F. 1942. Scientism and the study of science. *Economica* 9: 267–291.
Watkins, J. 1952. The principle of methodological individualism. *British Journal for the Philosophy of Science* 3: 186–189.
Weber, M. 1922. *Economy and society.* Vol. 1. New York: Bedminster Press, 1968.

Methodology

Lawrence A. Boland

The term 'methodology' refers to the study of methods, usually, the study of scientific method. For most of this century, the concept of a scientific method was that of a multi-stage recipe. In particular, it was the one alleged to be used by successful scientists for more than 300 years. The typical high-school science textbook started with a description of this allegedly successful, and thus proper, method of scientific investigation. It said, for example, that all science begins, as the first step, with the collection of data. The second step was the formation of an 'hypothesis' concerning the collected data and the third step was the formation of an experiment to test this hypothesis. If the hypothesis passed the test it was given the title of a 'theory'. If the theory survived the tests of other scientists, perhaps after years, then the theory was called a 'Law'. The key lesson that aspiring scientists were thereby taught was that if they were methodologically careful collecting their data and forming and testing their hypotheses, they were assured of success – and perhaps even rewarded with fame.

Today it will be difficult for anyone to find a high-school textbook with such an optimistically mechanical picture of scientific method and scientific discovery. Instead, today's science textbook begins with a statement of a few fundamental ideas that characterize the science in question. The textbook's contents are primarily demonstrations of a method of validation where it is shown that when used carefully these fundamental ideas can be employed to explain or describe all the phenomena of interest to the science in question. There is very little discussion of scientific method – at least not as a mechanical procedure. Today, being 'scientific' is no longer considered synonymous with being true but is more commonly considered synonymous with what might be called 'rational error avoidance'.

For most of the last 200 years a primary symbol of intellectualism (especially among economists) was the ability to display a thorough understanding or proper methods of scientific investigation. Since World War II, however, it appears that any overt expression of interest in methodology is considered a clear sign of weak-mindedness or premature senility. Nevertheless, there is on-going discussion of methodology over morning coffee in almost every economics department. But, the contents of such morning discussions or arguments are usually concerned with little more than the best way to state the commonly accepted method. Depending on the revealed preferences of the department, the accepted method is either

some form of Paul Samuelson's 'descriptivism' (see Samuelson 1947, 1963, 1965; see also Wong 1978) or some variant of Milton Friedman's 'instrumentalism' (see Friedman 1953; see also Boland 1979).

Descriptivism is the method where theories are not considered explanations but only better or worse analytical descriptions of observable phenomena. Instrumentalism goes further by saying that theories are only instruments used either to make predictions for the purposes of assisting economic policy-makers or to make practical measurements of the essential parameters of the real world. Each of these accepted methods is beyond question among its followers. Neither group accepts any reason to study methodology. Criticism of their method is considered a waste of time and further justification is considered unnecessary.

Given this atmosphere, why would anyone want to study methodology? In the 1980s, this closed-minded attitude is breaking down as indicated by the publication of several books about methodology in economics beginning with Mark Blaug's self-conscious 'appraisal' of how economists explain (Blaug 1980). Unfortunately, Blaug's book is still an attempt to reestablish methodology as the study of the *one* proper method. This begs two questions. Why should there by only one proper method? And, why should any methodologist's appraisal of the work of economists ever matter? Students of methodology seldom consider these questions even though they are central to the study of methodology of economics.

Economists still study methodology. For some the reason is to acquire an unassailable basis for a criticism of mainstream economics. For a few others the reason is to acquire an unassailable basis for the justification of mainstream economics. Unfortunately, both of these reasons presume that there is only one correct method. To avoid the limitations of the narrow view, many methodologists are turning to what Bruce Caldwell (1982) calls 'methodological pluralism'. On the basis of pluralism there is now a conceivable third reason to study methodology. Methodologists may wish to understand why mainstream economics is what it is without making judgements or criticisms. For later reference this new pluralistic approach will be called 'cognitive methodology'.

The overwhelming success of the physical sciences has given considerable support to the belief in the existence of a single, correct scientific method. If only there were an unambiguous method which if followed to the letter would give us perfect theories, then everyone would agree that it would be wise to follow that method in the development of economic analysis. The existence of such a perfect method is, however, a romantic dream. Nevertheless, it is always tempting for us to believe in such a method whenever facing frustrating theoretical or ideological opponents – particularly so when confronting theorists whose viewpoints are obviously mistaken.

The key question is: whenever mistakes are made in the development or validation of a theory, can they always be explained as failures to follow a proper method? If the answer is affirmative, then it might be believed that all failures are due to choosing the wrong method or due to using the correct method improperly. If this line of reasoning is accepted, almost every criticism of someone's views or theories could be seen to be a claim of methodological improprieties.

There are many so-called methodological disputes that are really exercises in ideological differences. For example, arguments between Marxists and mainstream economists are often irresolvably at cross-purposes. The resolvable disputes should be those between different mainstream economists. For the last twenty-five years the paradigm of such methodological disputes has been fostered by Samuelson's criticism of Friedman's instrumentalism (Samuelson 1963). Much of the continuing criticism of Friedman's view of methodology is really an objection to his opposition to government intervention. It is doubtful whether any other opponent of government intervention would express any objection to Friedman's instrumentalist methodology. Surprisingly, there is very little criticism of Samuelson's descriptivism – but this is easy to understand. Proponents of Friedman's methodological opinions argue that actions speak louder than words.

Almost all spectators of this paradigmatic methodological dispute miss the point that Friedman's view of methodology is itself primarily an attack on the 1930s analytical philosophy of 'Positivism'. Positivism (or 'Modernism' as Donald McCloskey (1983) calls it) was the view that theories are scientific whenever their assumptions are logically capable of being verified either introspectively or empirically (see Caldwell 1982). Friedman's so-called methodology, then, is not an alternative scientific method of developing a theory. It is only a critique of an old view that was based on the existence of a method which it followed would lead to positive results even though the method depends only on *a priori* analysis of the assumptions used to develop theories.

It is easy to show that many current views of methodology are really expressions of critiques of other views of methodologically. Friedman criticizes 1930s positivism (which itself was a criticism of 18th-century empiricism) and Samuelson criticizes Friedman's instrumentalism. It might be wondered who will come forth to criticize Samuelson's alleged methodology and thus become the target for the next methodological critique. For the present, however, it is clear that criticism is a most important aspect of these modern disputes.

Methodology has not always been an exercise in sophisticated criticism. There are plenty of apologetic essays on 'the scope and method of economics'. As explained by Blaug, during the 19th century there were straightforward expositions of economic methodology beginning with Nassau Senior and followed by John Stuart Mill and John Elliot Cairnes leading finally to the summary by John Neville Keynes. While these exposition did occasionally examine some problems with economic methodology, they were more often concerned with explaining how economists can successfully go about their work (see Blaug 1980, Part 2). In this century, Lionel Robbins (1932) seems to have been the last methodologist to try to explain why economists are so successful.

Given the progressive development of economics in the last fifty years one might expect that there would be more attempts to explain the apparent success of mainstream economics. Such is not the case. It is true that most principles textbooks do contain an introductory chapter that discusses methodological issues such as the distinction between 'positive' and 'normative' statements, the idea of a tautology, and various logical fallacies. But these chapters and discussions are just cosmetic touches and they fall far short of being the traditional expositions of the 'scope and methods' of economics.

It is difficult if not impossible to avoid taking a position with respect to methodology whenever making decisions which require information, knowledge or the formation of expectations. Methodology is pervasive regardless of whether it is fashionable to recognize methodology as a legitimate area for study by economists. If the currently popular opinions of economists trained since the early 1960s are to be believed, the study of methodology is dead or methodology is non-existent. This would be very misleading. Methodology lives but it is not easy to see anymore because it is embodied in the accepted hidden research agenda (see Boland 1982). Today the accepted methodology has to be inferred from the uniformities in actual practice of economists.

Since the 1960s there has been a uniform growth of one specific view of the 'correct method' or scientific investigation. This view dominates both the criteria of journal editors and decisions of curriculum committees of major economics departments. The evidence of this dominant view of methodology is the growth of mathematical formalism. While the journals in the early 1950s contained very little formal mathematics, by the late 1970s almost all leading journals were devoting most of their space to articles that were either completely concerned with the mathematical analysis of invented mathematical models or with methods of presenting economics ideas using mathematical formalism.

These observations are noted to illustrate that there has been a change in the view of the one method that many think should be used in economics. Cognitive methodology is interested in why economists choose to present or develop their theories the way they do. No appraisal is

necessary. While it might be argued that mathematics does not matter since it is only a language, it would be difficult to see how much an argument can explain why so much journal space has been devoted to proofs of mathematical theorems rather than to the explication of economic ideas and data. The question cognitive methodology asks is just what method when practised leads to an emphasis on mathematical ideas rather than on economics ideas and data? It cannot be the old science-textbook methodology since that view emphasized the quality of data and data collection.

The old textbook view of methodology was based on a 500-year-old theory of learning that was refuted 200 years ago. The old theory of learning claimed that one could learn from data alone and that all knowledge would be shown to follow logically from collected data. It was supposedly always possible to give a true explanation of the facts of the real world mechanically using only prior factual observations. David Hume convincingly argued that such a view of knowledge is self-contradictory or an infinite regress. For example, how do we know when a collection of data is a true representation of the world? Our knowledge of our knowledge must also depend only on data and this dependence would lead either to another challenge or to an inconsistency.

Hume's arguments have led many to say that the problem with the old view is its presumption that knowledge can be based solely on data without relying on any auxiliary theories about how to collect or interpret data. The now popular new view says that all data and observations are 'theory laden' which means simply that the use of auxiliary theories is unavoidable. There are no pure observation reports and observations alone are insufficient for scientific method. Some may still ask, what explains the success of science if it is not the careful collection of data?

According to the old science-textbook methodology, the 'gentleman scientist' in the privacy of his home laboratory can carefully collect data and rationally manipulate that data to generate true theories of observable phenomena. Since his method is claimed to be rational, his theories will be accepted by anyone who could follow rational arguments simply because his theories follow logically from the collected data. Rationality of an argument means that anyone using the same premises will reach the same conclusions and when all premises of the argument are true then all of the conclusions will be true. According to the old view of science, the only premises used are the observation reports of the data. Examination of the foundation of the old view reveals that it was based on a belief that anyone could collect objectively true observational facts and that there existed an unerringly adequate inductive logic.

The new view has as its foundation the recognition that such an unerring inductive logic is impossible. The absence of a reliable inductive logic means that objective data alone will never be enough to explain any phenomenon. We always need to make assumptions in our explanations and thus the acceptance of our explanations or descriptions must be based on the acceptance of our auxiliary theories.

The recognition of a necessary role for auxiliary theories leads to the view that if we wish to avoid an infinite regress and still be concerned with successful explanations (or descriptions) of the real world, then we must have some method to decide which auxiliary theories are acceptable. On what basis do we accept one auxiliary theory and reject another? Facts alone cannot be appealed to since any report of the facts is considered 'theory-laden' and thus would only beg questions about other auxiliary theories. What more can we demand without appealing to the indisputable facts that are claimed to be impossible? Well, all auxiliary theories must at least be logically valid. This raises a question of how the logical validity of a theory is assured. Few economists are trained in formal logic. But such training is unnecessary whenever the economist is willing and able to use readily available mathematical theorems that have been proven by competent mathematicians. If the economist's auxiliary theories satisfy the demands of the mathematics profession, then it might be said that no further justification is needed.

Some methodologists might wish to defend the growth of mathematical formalism in economics

on the grounds that today every scientific discipline such as physics or chemistry would be difficult to understand if we were to remove the mathematics from its methods of analysis. But many detractors have complained, the danger of the currently accepted view of the proper methodology is that too often the elegance of one's mathematical analysis is considered more important than the relevance of the results of such analysis. Despite the air of confrontation, it is actually possible to understand methodology without taking a stand between these two opposing views.

The distinction between the old and new views of methodology can be used here to claim that the growth of mathematical economics is a direct consequence of the shift from the old view to the new. In the old view scientists collect indisputable observational factors; in the new view scientists collect or create indisputable logically valid theorems which may or may not be about observable data. In the old view, properly collected facts would speak for themselves without the need of auxiliary theories of evidence; today, the only sure thing is the rationality of our arguments in favour of our theories. The new view that proving logical validity is more important than careful data collection can be seen to be the basis for the recent excessive emphasis of mathematics. However, we must be careful to recognize that the shift from pursuing indisputable empirical facts to pursuing indisputable logical facts is not fundamentally a major change in methodology. Nevertheless, the change is sufficient to explain the shift from concern for methods of collecting facts to methods of assuring the logical validity of our explanations or descriptions of the economy.

In cognitive methodology there is no need to judge the goodness or badness of mathematical methods in economic analysis. After 150 years of studying the methodology of economics it should be safe to say that there may be more than one correct methodology, or better, to admit that of the numerous candidates, the correct one may depend on what we want to do (see Boland 1982). The methodologist today must keep an open mind. If the economist is interested in dealing with very short-run practical problems, Friedman's instrumentalism might be appropriate. However, if we are interested in determining the formal similarities between various parts of economic theory, application of Samuelson's descriptivism could be a better guide. Even though there may not be an adequate inductive logic, the old science textbook methodology may still be a useful starting point. And, if we are interested in developing elegant models that might impress mathematicians, then we may find support in the new science textbooks' methodology.

Obviously the study of methodology is important for historians of economic thought. If methodology is an important basis in explanations of economists, it is also an important ingredient in the explanation of all decision-makers. Given a plurality of methodological views, we can no longer assume that the individual whose behaviour theorists explain shares the same methodology as the one used by the theorists. That is, we can no longer assume that there is just one rational method. Methodology today is then becoming the study of how different individuals deal differently with information or data in the process of making their decisions.

See Also

- ▶ Epistemological Issues in Economics
- ▶ Methodenstreit
- ▶ Models and Theory
- ▶ Philosophy and Economics
- ▶ Positive Economics
- ▶ Positivism
- ▶ Rhetoric of Economics

Bibliography

Blaug, M. 1980. *The methodology of economics: Or how economists explain*. Cambridge: Cambridge University Press.

Boland, L. 1979. A critique of Friedman's critics. *Journal of Economic Literature* 17: 503–522.

Boland, L. 1982. *The foundations of economic method*. London: George Allen & Unwin.

Caldwell, B. 1982. *Beyond positivism: Economic methodology in the twentieth century*. London: George Allen & Unwin.
Friedman, M. 1953. On the methodology of positive economics. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.
McCloskey, D. 1983. The rhetoric of economics. *Journal of Economic Literature* 21: 481–517.
Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P. 1963. Problems of methodology: Discussion. *American Economic Review, Papers and Proceedings* 53: 231–236.
Samuelson, P. 1965. Professor Samuelson on theory and realism: Reply. *American Economic Review* 55: 1164–1172.
Wong, S. 1978. *The foundations of Paul Samuelson's revealed preference theory*. London: Routledge & Kegan Paul.

Methodology of Economics

Roger E. Backhouse

Abstract

The methodology of economics concerns the principles underlying economic argumentation. Though systematic reflections on the subject go back at least to the 19th century, the methodology of economics emerged as a recognizable field, at the boundaries of economics, philosophy and science studies, only in the 1980s. This article outlines the way the field developed and its current position.

Keywords

Assumptions controversy; Behavioural economics; Cairnes, J.E.; Causality; Conventionalism; Data mining; Economics, definition of; Experimental economics; Explanation; Falsificationism; Friedman, M.; German Historical School; Heterodox economics; History of economic thought; Ideology; Individualism versus holism; Instrumentalism; Jevons, W.S.; Keynes, J.N.; Koopmans, T.C.; Kuhn, T.; Lakatos, I.; Leslie, T.E.C.; Measurement theory; Menger, C.; Mill, J. S.; Models; Operationalism; Paradigms; Philosophy and economics; Pluralism in economics; Popper, K.; Positive economics; Positivism; Postmodernism; Robbins, L.C.; Stylized facts; Theory appraisal

JEL Classifications

B4

Since the 1970s, the methodology of economics has developed from a series of reflections by practising economists on the methods employed in their field, to a field at the boundaries of economics and philosophy (and to a lesser extent sociology). After an initial focus on falsificationism, the range of issues pursued has considerably broadened.

In the social sciences, which include economics, the term 'methodology' is used with two different meanings. When an article or thesis contains a section called 'methodology' in which the author explains how a piece of research was conducted the word is used as a synonym for 'method'. In the literature on 'the methodology of economics', on the other hand, it is used as a label for enquiries into the principles underlying economic reasoning. This article is concerned only with the second of these two meanings.

The methodology of economics is inevitably an interdisciplinary activity. Economists may analyse their own reasoning using ideas drawn from philosophy, sociology, linguistics or discourse analysis, or they can draw simply on their own experience as practising economists. For philosophers, enquiries into economic methodology are part of philosophy – a branch of the philosophy of science or, if the word science is thought inappropriate, of knowledge. Economic methodology is thus largely covered by the article on philosophy and economics. The two are, however not synonymous, for the latter covers decision theory, rational choice and ethics, fields not

traditionally thought of as economic methodology. That article traces the interrelations between these disciplines from the 19th century to the present. However, though this overlaps with the story of economic methodology, the two are not synonymous.

The explicit study of methodology has always had a mixed reputation within economics. Most of the time, economists simply get on with their work, reflecting on specific methodological problems as and when they arise, refraining from more general speculation. They are suspicious of general theories about how to practise economics (or any other subject for that matter), especially when such theories are written by those who do not themselves engage in the research they are analysing. Against this there are those who believe that methodological reflection by those who are more distant from practice, whether they are trained as economists or philosophers, even if it does not tell economists how to do their work better, can provide a valuable perspective on what economists do that would be otherwise be missed. When it comes to publication, some, even if they find methodological argument valuable, hold that it should not have a place within economics journals as it is not economics, but writing about economics. A further reason for scepticism is that methodological arguments are frequently used by non-economists and heterodox economists to show that certain economic theories cannot possibly be right: it is held that, rather than speculate on methodology, those who believe this would do better if they showed by example, how things could be done better. This attitude has a parallel in divisions within the field of economic methodology between those who believe that the task of methodology is primarily to understand what economists do (a stance that does not imply an absence of criticism, even if the methodologist deliberately refrains from telling economists what to do) and those who use methodological arguments to argue for heterodox positions within economics. These two categories overlap significantly, but this divide nevertheless reflects important tensions within the field.

Historical Background

The 19th century was an age when disciplinary boundaries began to be established. Given the extremely high regard in which John Stuart Mill was held by contemporaries, both as a philosopher and as an economist, it would be rash to classify him according to modern disciplinary categories. His *Logic* (1843) was a standard textbook in the philosophy of science and his *Essays on Some Unsettled Questions of Political Economy* (1844) were an influential statement of economic methodology. Methodological arguments among economists were frequent (see, for example, the work collected in Smyth 1962; Backhouse 1997) and were primarily by economists using methodological arguments to criticize positions with which they disagreed. Cliffe Leslie and John Elliott Cairnes are good examples. Both established reputations for their work in economics itself, but wrote extensively on methodology, Leslie in a series of essays (1879) and Cairnes in *The Character and Logical Method of Political Economy* (1857). William Stanley Jevons made a methodological case for a particular way of practising economics in his *Theory of Political Economy* (1871) but in addition to being a leading economist was also the author of *The Principles of Science* (1873), a major textbook in the philosophy of science. If anyone should be classified as a professional methodologist in this period, it is John Neville Keynes, author of a textbook in formal logic, but whose *The Scope and Method of Political Economy* (1890) was his major work. Even if some considered it a worthy book, but one that students did not in practice need to bother with, it played a role in establishing the Marshallian consensus within British economics and preventing the methodological dispute between the Carl Menger (1883) and the German Historical School from dividing the British profession in the same way as it divided German-speaking economists. There were methodological disputes over historicism in British economics, but they were nowhere nearly as divisive as the German.

The tradition of economists writing article and occasionally book-length reflections on

methodology continued through the 20th century, and was linked to disputes over the direction in which economics should be moving. In the first half of that century, the most influential such work was undoubtedly Lionel Robbins's *An Essay on the Nature and Significance of Economic science* (1932), which helped define modern welfare economics and, more broadly to redefine the subject, though this took much longer than is commonly believed (Backhouse and Medema 2007). The 1930s saw a profusion of articles and books on methodology, many of which discussed Robbins's *Essay*. However, what we find is a literature that, though containing much that was perceptive, can be seen as a series of comparatively isolated works in which trends are hard to identify.

After the Second World War, this pattern continued, though the literature became more focused, due to the way economic theory was developing. There was also, in the background, the emergence of what came to be known as the 'received view' in the philosophy of science and the work of Karl Popper, though lags in translation meant that this permeated the economic literature only gradually. Several of the leading economists wrote on methodology, their work being given focus, despite their varied perspectives, by their concern with models and the role of assumptions. The most influential work was Milton Friedman's 'The Methodology of Positive Economics' (1953) with its provocative thesis that it was actually desirable for the assumptions of a theory to be unrealistic. Good theories are ones that pick out the relevant features of reality, using a sparse set of assumptions to explain a wide range of phenomena, which means that they must be descriptively unrealistic. Tjalling Koopmans included an essay defending unrealistic models on quite different grounds in his *Three Essays on the State of Economic Science* (1957): unrealistic models should be seen not in isolation but as part of a series of models – they are prototypes of subsequent models that will be more realistic. These, and above all Friedman's essay, provoked a significant literature. The question of testability linked this with issues being discussed in the philosophy of science in a way that was not generally true of the period before the Second World War (an exception was Hutchison 1938).

Methodology of Economics as a Field

The emergence of the methodology of economics as a recognizable field within economics, into which this earlier literature could (retrospectively) be incorporated, is best dated to the appearance of Mark Blaug's *The Methodology of Economics* (1980). This was not the first textbook on economic methodology but it served to define a field in a way that previous textbooks (for example, Stewart 1979) had not. It offered a survey of what economists needed to know about the philosophy of science and a series of case studies in economics. The theme of the book was the importance of falsificationism, as found in the work of Karl Popper and Imre Lakatos. He offered a typically robust conclusion:

> [T]he ultimate question we can and must pose about any research program is the one made familiar by Popper: what events, if they materialized, would lead us to reject that program? A program that cannot meet that question has fallen short of the highest standards that scientific knowledge can attain. (Blaug 1980, p. 264)

The common practice among economists was what he called 'innocuous falsificationism': to preach falsificationism but not to practise it. The theme of his case studies was that the subject had made progress when economists had sought to test theories, even when such testing had, as in the examples of human capital theory or monetarism, been inconclusive.

Blaug's book was important. His thesis offered a challenge, both to those who felt that there must be reasons why economists approached their subject in ways that, according to Blaug, were fundamentally flawed, and to those who were concerned about the philosophical coherence of falsificationism. This defined a research agenda. His approach to methodology also pointed to ways in which it could be combined with the history of economic thought. Popperian methodology, especially in its Lakatosian variant, with its

focus on progress, provided a criterion that could be used to assess the history of economics, an approach already explored in Latsis (1976). For the first time, economic methodology came to be linked with the history of economic thought.

A further stimulus to economic methodology came from heterodox economics. Movements such as radical economics, Post Keynesian economics and Austrian economics, though their proponents might construct longer histories, originated in the late 1960s and early 1970s, and were characterized by methodological critiques of the way they saw economic enquires being undertaken. Their interest in both methodology and the history of economics brought tensions: their interest in the subject was welcomed but this was associated with concerns that their ideological commitments might cause a problem for the field (in his article on history of economic thought, Goodwin hints at similar concerns).

The result of this activity was the emergence of an identifiable field of economic methodology, defined not simply by textbooks but by a community of specialists engaging with each other as well as economists and philosophers who chose to explore the subject. This blend of economics and philosophy was reflected in the specialist journals, those publishing articles in English including *Economics and Philosophy* (established 1985), *Journal of Economic Methodology* (1994), *Research in the History of Economic Thought and Methodology* (1983), and in anthologies such as Caldwell (1993) or Hausman (1994) (the former contains a useful list of previously anthologized articles). There are also foreign-language journals, of which *Revue de philosophie économique* has begun publication in English. The difference between the *Journal of Economic Methodology* and *Economics and Philosophy* illustrates the point made earlier that, though there is substantial overlap, economic methodology is not synonymous with philosophy and economics.

The emergence of the field of economic methodology was centred on the philosophy of Popper and Lakatos. By 1990, these approaches had ceased to be dominant. Though some (for example, Blaug, Lawrence Boland and Terence Hutchison) continued to defend it, falsificationism was generally seen as too restrictive a criterion against which to appraise economic theories: the methodologies of Popper and Lakatos had technical problems and there were good reasons why economists behaved differently. The most significant problem arises from the fact that theories are almost never testable on their own, creating problems with what it means for a theory to be falsifiable (see falsificationism). Lakatosian methodology raised questions concerning the definition of a research programme and the meaning of novelty (see paradigms). Corroboration seemed more important than either Popper or Lakatos admitted.

The most influential alternative was articulated by Daniel Hausman's *The Inexact and Separate Science of Economics* (1992). Dismissive of Popper and Lakatos, this book opened up other themes, such as ways of thinking about economic models, but its significance was engaging in what Hausman elsewhere summarized as 'empirical philosophy of science'. This involved exploring in detail economists' practices, his main example being the way economists had responded to the phenomenon of preference reversals. Hausman defended the view that economics was, as Mill had expressed it in the mid-19th century, an inexact science, but he was more critical of the motivation, implicit in economists' responses to experimental evidence, to keep their science separate from any dependence on philosophy. This conclusion may, at least in part, have been rendered out of date by the rise of behavioural economics, but its significance lay in the method of starting from economics, drawing out methodological conclusions (as opposed to the method) characteristic of the Popperian era, and of applying the methodology developed in the context of natural science to economics.

The problems involved in defining Lakatosian research programmes were widely considered to render the concept a rather blunt tool for analysing economics. Instead, the trend was towards analysing problems that arose in particular fields of economics. The rise of experimental economics

raised new methodological questions about experimental procedures and the transferability of experimental results to behaviour out of the laboratory. Econometric practices raised issues not covered in traditional methodology such as the significance of data mining, the meaning of causality and how measurement of economic quantities should be understood. Disagreements over the relation between macro and microeconomics raised questions about individualism and the meaning of aggregate analysis, many of which were familiar to the philosophy of social science. Economists had come, almost universally, to argue in terms of models, raising the question of what was going on in the process of economic modelling. Postmodernism failed to have anything like the effect that it had in some other social sciences, but some postmodern ideas were explored. Older questions, such as conventionalism, instrumentalism, positivism and falsification, all concerned with questions of theory appraisal, remained, but they received proportionately much less attention.

The Methodology of Economics Today

The methodology of economics has remained a field with very elastic boundaries. There is a substantial literature in which specialists in the field engage with each other, but perhaps more than most fields in economics, it is one where outsiders, who engage in varying degrees with this literature, have much to say. These outsiders include philosophers, economists specializing in other fields and other social scientists. This results in a variety of perspectives, one of the main divides concerning the extent to which writers see methodology as aimed at understanding economists' practices and the extent to which they seek to criticize those practices. Sometimes these aims overlap, but sometimes they do not, those concerned with explication considering that others do not take economics sufficiently seriously, and those concerned with criticism considering that the others are too defensive about what they find within the discipline.

The result is that it has become increasingly difficult to find a framework within which to offer a coherent survey of the field. The most comprehensive recent attempt is Wade Hands's *Reflection without Rules* (2001), which he describes as an 'interpretive survey' of the economic methodology. Part of the difficulty with such a task is, as Hands (2001, p. ix) recognized, that any such survey is aiming at a moving target. That would be true in any field; however, a further difficulty is that much work on methodology cuts across the philosophical categories he employs to structure his survey and hence provide the basis for his interpretation. His starting point (as for Blaug 1980; Caldwell 1982), was the breakdown of the received view within the philosophy of science, after which he identified a series of turns – the naturalistic, the sociological, the pragmatic and the economic – as well as attempts to develop the Popperian, Millian and other traditions. From the point of view of elucidating the philosophical foundations of economic methodology, this is a successful strategy. However, this framework does not shed much light on some practice-based methodological work. For example Hoover's (2001a, b) work on macroeconomics and causality could be fitted into its framework but, despite its deep philosophical engagement, it is not clear that it is helpful to do so. The problems with work by reflective practitioners, such as Reder (1999) or Goldfarb (1997), simply does not fit at all.

Economic methodology has become a very active field that tackles a range of questions that is much broader than was the case in, say, 1950. In part this reflects the broadening that has taken place within economics, which has meant that methodologists have faced the challenge of broadening their focus to encompass developments as varied as experimental economics and time-series econometric methods. There has also been a shift away from abstract questions of theory appraisal towards understanding the variety of practices found in economics, the aim being to work out a philosophical framework that is appropriate to economics rather than simply applying one derived from consideration of natural science. Finally, economic methodology has turned not

only to philosophy of science, as that term has traditionally been understood, but also to disciplines such as sociology, linguistics and science studies. This plethora of new developments suggests that it may change as much in the next quarter century as it has done in the past.

See Also

- Assumptions Controversy
- Causality in Economics and Econometrics
- Conventionalism
- Data Mining
- Economics, Definition of
- Experimental Methods in Economics
- Explanation
- Falsificationism
- Historical School, German
- History of Economic Thought
- Individualism Versus Holism
- Instrumentalism and Operationalism
- Measurement
- Models
- Paradigms
- Philosophy and Economics
- Pluralism in Economics
- Positive Economics
- Positivism
- Postmodernism
- Stylized Facts
- Theory Appraisal

Bibliography

Backhouse, R.E. 1997. *The methodology of economics: Nineteenth-century British contributions*, 6 vols. London and Bristol: Routledge & Thoemmes Press.

Backhouse, R.E., and S.G. Medema. 2007. Defining economics: Robbins's essay in theory and practice. Working paper. http://ssrn.com/abstract=969994. Accessed 19 May 2007.

Blaug, M. 1980. *The methodology of economics: How economists explain*, 2nd ed. Cambridge: Cambridge University Press, 1992.

Cairnes, J.E. 1857. *The character and logical method of political economy*. London: Macmillan, 1888.

Caldwell, B.J. 1982. *Beyond positivism: Economic methodology in the twentieth century*. London: Allen & Unwin.

Caldwell, B.J. 1993. *The philosophy and methodology of economics*, 3 vols. Cheltenham: Edward Elgar.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Goldfarb, R. 1997. Now you see it, now you don't: Emerging contrary results in economics. *Journal of Economic Methodology* 4: 221–244.

Hands, D.W. 2001. *Reflection without rules: Economic methodology and contemporary science theory*. Cambridge: Cambridge University Press.

Hausman, D.M. 1992. *The inexact and separate science of economics*. Cambridge: Cambridge University Press.

Hausman, D.M. 1994. *The philosophy of economics: An anthology*. Cambridge: Cambridge University Press.

Hoover, K.D. 2001a. *The methodology of empirical macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K.D. 2001b. *Causality in economics*. Cambridge: Cambridge University Press.

Hutchison, T.W. 1938. *The significance and basic postulates of economic theory*. London: Macmillan.

Jevons, W.S. 1873. *The principles of science: A treatise on logic and scientific method*, 2nd edn. London: Macmillan, 1877.

Jevons, W.S. 1871. *The theory of political economy*, 2nd edn. London: Macmillan, 1879.

Keynes, J.N. 1890. *The scope and method of political economy*. London: Macmillan, 1917.

Koopmans, T.C. 1957. *Three essays on the state of economic science*. New York: McGraw Hill.

Latsis, S.J. 1976. *Method and appraisal in economics*. Cambridge: Cambridge University Press.

Leslie, T.E.C. 1879. *Essays on political and moral philosophy*. Dublin: Hodges, Foster & Figgis.

Menger, C. 1883. Investigations into the method of the social sciences with special reference to economics. Trans. F.J. Nock. New York: New York University Press, 1985.

Mill, J.S. 1843. *A system of logic. The collected works of John Stuart Mill*, vols 7–8, ed. J.M. Robson. Toronto: University of Toronto Press, 1974.

Mill, J.S. 1844. Essays on some unsettled questions of political economy. In *The collected works of John Stuart Mill*, vol. 4, ed. J.M. Robson. Toronto: University of Toronto Press, 1967.

Reder, M.W. 1999. *Economics: The culture of a controversial science*. Chicago: University of Chicago Press.

Robbins, L.C. 1932. An essay on the nature and significance of economic science. London: Macmillan, 1935.

Smyth, R.L. 1962. *Essays in economic method: Selected papers read to Section F of the British Association for the advancement of science, 1860–1913*. London: Duckworth.

Stewart, I. 1979. *Reasoning and method in economics*. New York: McGraw Hill.

Metzler, Lloyd Appleton (1913–1980)

George Horwich and John Pomery

Keywords

Dynamic stability; Flexible exchange rates; Gross substitutes; Income determination; Inventory cycles; Inventory investment; Laursen–Metzler effect; Matrix theory; Metzler matrix; Metzler paradox; Metzler, L. A.; Optimal tariffs; Perfect stability; Propensity to consume; Purchasing power parity; Real balances; Stability; Tariffs; Terms of trade; Transfer problem; Wealth effect

JEL Classifications

B31

Metzler was born in Lost Springs, Kansas and took AB and MBA degrees at the University of Kansas. He was one of a long line of brilliant students that John Ise sent to Harvard, where Metzler arrived in 1937. He served as an instructor and tutor, receiving his Ph.D. in 1942. From 1943 to 1946 he held a number of positions with government agencies in wartime Washington, including the Office of Strategic Services, several economic policy and planning commissions, and, from 1944 to 1946, with the research staff of the Board of Governors of the Federal Reserve System. From 1946 to 1947 he was a member of the economics department of Yale University. In 1947 he joined the department at the University of Chicago, where he remained until his death. His health declined in the early 1950s; removal of a brain tumour in 1952 left him with a markedly reduced energy level. He continued to teach and produced an occasional paper for the next 20 years.

Metzler's *Collected Papers*, most of them written between 1941 and 1951, were published by Harvard University Press in 1973. A Festschrift, *Trade, Stability, and Macroeconomics*, co-edited by his fellow student Paul Samuelson and one of his own students, George Horwich, was published by Academic Press in 1974.

Metzler's contribution to the business cycle literature centred on his integration of inventories into a dynamic model of income determination. Employing the income/expenditure multiplier/accelerator framework pioneered by Robertson, Keynes, Hicks, Lundberg, Samuelson and others, Metzler added a rigorous formulation of inventory behaviour against a backdrop of production lags and endogenously determined anticipation of sales.

His initial assumption, supported by his later empirical investigation (1948), was that the response of output to sales receipts was the longest of the three basic lags in the circular flow of income, of which the other two were the lag between the receipt of income and consumption spending and the lag between output and the distribution of earnings to factors of production. Featuring the output/sales lag, his classic study, 'The Nature and Stability of Inventory Cycles' (1941), demonstrated that any disturbance, such as an autonomous increase in investment, tended to produce cycles about the new level of income provided that the marginal propensity to consume is less than unity. The cycles are damped if businesses demand a constant level of inventories and expect sales in the current period to be unchanged at the level of the preceding period. Explosive cycles may occur if anticipated sales change when actual sales change and if firms try to maintain a constant ratio of inventories to anticipated sales.

If inventory demand varies with anticipated sales, a reduction in inventories below desired levels during an expansionary phase of the business cycle raises the demand for inventories and hence reinforces the rise in income and expenditures. The increases gradually taper off, causing a fall in planned inventory investment. Income and sales therefore fall and inventories rise about desired levels. The demand for inventories drops further and reinforces the decline in income. The model thus gives rise to predictions about the relative timing of movements in income, sales,

and inventories, and of response coefficients for which the cyclical process will converge to a new equilibrium.

Investigations in the 30 years following the 1941 paper, including several of his own (1946, 1947b, 1973e), tended to support Metzler's initial formulation. The basic model was also enriched by the addition of monetary factors and the rate of interest, the price level, and a disaggregation of inventories into finished goods and goods in process (Zarnowitz 1985, pp. 541–2).

Metzler's contribution to macro-monetary theory came from a single influential paper, 'Wealth, Saving, and the Rate of Interest' (1951b). Metzler wrote in the wake of the great debates between Keynes and his critics (Haberler 1941; Pigou 1943, 1947; Scitovszky 1941), who had invoked a positive relation between real cash balances and expenditures on goods and services as the basis of achieving a stable macro equilibrium. Metzler, taking a broader view of wealth as including both real balances and financial claims to the capital stock, argued that the implied inverse wealth/saving relation introduced a monetary element in the determination of the interest rate. Whereas money was traditionally without any lasting influence on the real side of the 'classical' model, the presence of the wealth/saving relation meant that the exchange of money and securities (the prototype of which is an interest- bearing equity claim) between the central bank and the private sector altered the latter's perceived wealth total and hence its rate of saving and the equilibrium rate of interest.

In Metzler's analysis, an open-market purchase, for example, raises cash balances and removes securities from private portfolios. In the resulting adjustment process, real balances are reduced by the rise of the general price level, but securities are not restored. The consequent net wealth loss stimulates a greater rate of saving and a lower rate of interest in the post-purchase equilibrium.

Most commentators disputed Metzler's claim that in the classical tradition, monetary changes tended not to affect the equilibrium rate of interest (Haberler 1952, p. 245; Patinkin 1956, pp. 260–1). On the other hand, the early critics would probably have agreed that Metzler was the first to articulate a specific influence on the interest rate that sprang from open-market operations in a model containing the wealth/saving relation.

Later critics, notably Robert Mundell (1960), questioned the *direction* of that influence on the interest rate. Metzler had been careful to account for the disposition of the earnings on the securities acquired by the central bank in the open-market purchase. In order to prevent private disposable income from falling continuously as these earnings are received by the bank in the future, the fiscal authority is assumed to reduce taxes by an equal amount (Metzler 1951b, p. 109n). Mundell pointed out that if the taxes reduced are those on capitalizable income, the securities sold to the central bank are exactly replaced by an upward revaluation of remaining privately held securities. The wealth effect of the operations, taking account of the inflation-induced loss of real balances, is a 'wash'. A reduction in taxes on capital, however, also raises the net return to capital and hence the investment demand schedule. In the new equilibrium, the rate of interest is higher.

Metzler (1973d, pp. 354–62) replied that capitalizable federal taxes are at most 30 per cent of total federal taxes. Proportionately, 70 per cent of a tax cut falls on non-capitalizable personal income. Only a small part of the value of securities sold to the bank is thus recovered by any likely tax cut, and the operation's wealth effects remain predominantly as Metzler described them.

David McCord Wright (1952) pointed out that the lower equilibrium interest rate generated by Metzler's open-market purchase will itself promote a more rapid investment rate, offsetting thereby the community's loss of securities and real wealth due to the purchase. Metzler (1952) objected that such an offset fell outside the limited time frame that his analysis and macro-theory generally were properly concerned with.

George Horwich (1962, 1964) argued that offsetting wealth changes due to forced saving tend to characterize the very process by which the new equilibrium interest rate is reached. While Wright

saw a long-run wealth offset spurred by the wealth/saving-induced lower rate of interest, Horwich questioned whether the short- run adjustment creating a reduced equilibrium interest rate was viable. His characterization of the adjustment process (influenced by Metzler's account of the securities markets underlying the model) emphasized the equilibrating role of new or flow security supply and demand originating in new investment and saving, respectively. The excess of investment over saving created by the operation is thus, in its financial counterpart, an excess supply of new securities that directly moves the interest rate towards its new equilibrium and funds additional investment spending. If, through forced saving, the excess investment is realized in additional real capital, the excess new securities tend to replace those sold to the bank. The process of reaching any post-operation equilibrium is thus one in which additional security issues and increments of capital stock are necessarily involved and tend, more or less, to maintain the pre-operation level of wealth.

Niehans (1978) questioned Metzler's specification of the capital stock as a determinant of saving, while real balances, which are at desired levels in equilibrium, are not so specified. Both wealth components, Niehans argued (1978, pp. 91–2), should be at desired (optimum) levels in equilibrium and should not appear in individual demand functions. He saw the main contribution of the 'wealth' article in its elegant formulation of the neoclassical synthesis (see also Haberler 1952, p. 246 for this viewpoint), in the distinction it made in the differing impacts of the various types of monetary change, and in its broad influence on the methodology employed by the major monetary writers of the next quarter century.

At least half of Metzler's papers are related to the field of international trade. He is probably best known for papers on tariff theory, international macroeconomics, and the transfer problem, but other work includes a lucid survey (1949a), a discussion of difficulties of applying purchasing power parity in post- Second World War exchange rate realignments (1947a), and a discussion of the views of Frank Graham (1950a).

In tariff theory, Metzler's contribution has largely been summarized in the statement that, in a two-country, two-good world, a tariff can fail to be protective in the sense that it can lead to a reduction in the domestic relative price of the import- competing good. This is the so-called 'Metzler paradox'. It is ironic that Metzler himself described this result as well known, and viewed his own contribution as analyzing the implications for income distribution of such a non-protective tariff (1949b, p. 10). Metzler's papers on this topic were apparently motivated by pronouncements of Australian economists, and the two papers (1949b, c) include discussion of alternative assumptions about expenditure of tariff revenue – by government (non-dutiable) or the private sector (dutiable), with various marginal propensities to consume, and from a zero or non-zero initial tariff. However the cleanest and best-known result comes with zero initial tariff and with tariff revenue implicitly going to the private sector as an increase in disposable income. Metzler shows that the domestic relative price will not change if the elasticity of import demand in the foreign country is equal to 1 minus the marginal propensity to import in the home (tariff-levying) country.

This foregoing result is easily understood by considering the world market for the home importable. The imposition of a trade tax, at constant domestic relative price, will imply a vertical shift of the foreign export supply curve (as in an elementary tax-incidence problem), since foreign export supply is a function of world relative price. If the foreign offer curve is elastic, this implies decreased export supply by the foreign country at the fixed domestic price, but if the foreign offer curve is inelastic (implying that the export supply curve is backward-bending), then the result is increased export supply at a given domestic price. On the demand side, a fixed domestic price implies a lower world relative price of the home importable. Thus the home country is better off, via the improvement in its terms of trade. If the importable is a normal good, the improved real income in the home country implies a rightward shift of the home import demand curve. In a Walrasian-stable market, the domestic price falls if and only if the shifts in home import demand

and foreign export supply combine to yield excess supply at the initial domestic price. Metzler's condition is a requirement that the shifts in the two curves exactly offset each other. Thus, if the importable good is normal at home, an inelastic foreign offer curve is necessary, but not sufficient, for the Metzler paradox.

The subsequent literature has enshrined this simplest version of the non- protective tariff (despite at least one attempt to refute the result). It is now understood that, given the normality assumption, the Metzler paradox is inconsistent with the home country levying the so-called optimal tariff; thus the 'paradox' is just one of many possible consequences of a second-best situation (from a myopic national viewpoint).

Another result to bear Metzler's name is the Laursen–Metzler effect, in honour of the joint authors (1950). Laursen and Metzler were concerned with integrating models of flexible exchange rates which focused either on income–expenditure effects or on terms of trade effects, but not on both simultaneously. They posited a channel from devaluation through the terms of trade and onto expenditure; specifically, a deterioration in the terms of trade was assumed to lead to increased expenditure with given nominal income. This was then applied to discussion of the extent of insulation via flexible exchange rates, and of the 'acceptability' of exchange rate changes in certain policy scenarios. The Laursen–Metzler effect has been integrated into the literature, although it was eclipsed in periods where there was extreme emphasis on flexible product prices (thus weakening the link from exchange rate changes to changes in the terms of trade) and although there is some question as to the sign of the effect. In a period of current-account and government- budget imbalances, there has been some emphasis on real intertemporal models of trade. With more sophisticated models of simultaneous intertemporal and intratemporal optimization than were available to Laursen and Metzler, a deeper understanding of the link between intratemporal terms of trade and current expenditure is now possible.

Many of Metzler's international papers involved the transfer problem; see Metzler (1942, 1951a, c, 1973b, c). Metzler's focus was on endogenous income and expenditure effects, holding prices and interest rates fixed. In later papers, the analysis is tied to Metzler's contributions to the applications of matrix theory to economics. In the transfer problem, since the initial transfer is a pure redistribution, the analytic question concerns what changes in endogenous variables are required to re-establish equilibrium. The pure trade literature has emphasized the endogenous adjustment of the terms of trade in real, flexible-price models – including a somewhat incestuous literature, mainly involving Samuelson and Jones, on the likelihood of orthodox or anti-orthodox bias. More recently, the possibility of 'paradox' in a multi-country setting has been the central topic. Metzler's assumption of constant prices led to analysis of the impact on trade balances and income at home and abroad, with discussion of the role of stability conditions nationally and globally and of the relevant roles for alternative income concepts in the presence of imported inputs for production. Chapter 4 of the *Collected Papers* (1973c) comes closest to linking up to the orthodox theory. While Metzler was one of many contributors to the transfer literature, his strong Keynesian perspective may have limited the long-term importance of his contribution more on this topic than on those mentioned earlier.

In the field of mathematical economics, Metzler has been honoured by having a matrix named after him. The central paper is perhaps Metzler (1945), but see also Metzler (1950b, 1951a, c). The Metzler matrix is a square matrix with positive diagonal elements, negative off-diagonal elements, positive principal minors and determinant, and a positive inverse matrix. Metzler investigated this class of matrices in the context of market stability (1945) and comparative statics (1950b). The stability analysis linked the Hicksian concept of market stability, which can be interpreted as essentially static, and Samuelson's explicitly dynamic approach to stability.

Metzler showed that if multiple markets are stable for any (relative) speeds of adjustment, then they must satisfy Hicks's concept of perfect stability. Perfect stability says that a fall in price in any single market creates excess demand in that

market – after any subset of other prices is adjusted to clear the 'own' markets – and all other prices are held fixed. Metzler's proof revolved about the alternating sign of principal minors of a matrix of partial derivatives of excess demands with respect to prices, the negative of this matrix leading to a Metzler matrix. Metzler also showed, by counterexample, that Hicksian perfect stability does not imply Samuelsonian dynamic stability. Another Metzler result showed that in the presence of gross substitutability, dynamic stability and perfect stability are equivalent. Gross substitutability guarantees that the matrix of partial derivatives has negative diagonal terms and positive off-diagonal terms, so that its negative has the sign pattern of a Metzler matrix. The intuition of the gross substitute case is that the impact of a change in 'own price' on excess demand for a good exceeds the aggregate impact of all 'other price' changes; thus, in a sense, the system generalizes the intuition of single-market stability analysis. While cross-effects can exist, the own effects dominate in each market. Metzler applied this theory to the comparative statics of fixed-coefficient regional models, multicountry income transfers, and taxes and subsidies in fixed-coefficient models. As is better understood after integrative work on matrices with dominant diagonals (McKenzie 1960) and on P-matrices (Gale and Nikaido 1965), strong results in 'square' – that is, $n \times n$ systems – usually require strong assumptions closely related to the existence of appropriate Metzler matrices. While there were many other contributors in this area, for example Hawkins and Simon, and while the majority of key results were already known to mathematicians, Metzler's work provided a crucial synthesis of stability literature and an important step in the evolution of matrix theory as applied in economics.

Selected Works

1941. The nature and stability of inventory cycles. *Review of Economics and Statistics* 23(3):113–129.

1942. The transfer problem reconsidered. *Journal of Political Economy* 50:397–414.

1945. Stability of multiple markets: The Hicks conditions. *Econometrica* 13(4): 277–292.

1946. Business cycles and the theory of employment. *American Economic Review* 36:278–291.

1947a. Exchange rates and the international monetary fund. In *International monetary policies*, ed. L.A. Metzler, R. Triffin, and G. Haberler. Postwar economic studies no. 7. Washington, DC: Board of Governors of the Federal Reserve System.

1947b. Factors governing the length of inventory cycles. *Review of Economics and Statistics* 29(1):1–15.

1948. Three lags in the circular flow of income. In *Income, employment, and public policy: Essays in honor of Alvin H. Hansen*, ed. L.A. Metzler. New York: W.W. Norton.

1949a. The theory of international trade. In *A survey of contemporary economics*, ed. H.S. Ellis. Philadelphia: Blakiston.

1949b. Tariffs, the terms of trade, and the distribution of national income. *Journal of Political Economy* 57:1–29.

1949c. Tariffs, international demand, and domestic prices. *Journal of Political Economy* 57:345–51.

1950a. Graham's theory of international values. *American Economic Review* 40:301–322.

1950b. A multiple-region theory of income and trade. *Econometrica* 18(4):329–354.

1950c. (With S. Laursen). Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32(4):281–299.

1951a. A multiple-country theory of income transfers. *Journal of Political Economy* 59:14–29.

1951b. Wealth, saving, and the rate of interest. *Journal of Political Economy* 59:93–116.

1951c. Taxes and subsidies in Leontief's input–output model. *Quarterly Journal of Economics* 65:433–438.

1952. A reply. *Journal of Political Economy* 60:249–252.

1973a. *Collected papers*. Cambridge, MA: Harvard University Press.
1973b. Imported raw materials, the transfer problem, and the concepts of income. In Metzler (1973a).
1973c. Flexible exchange rates, the transfer problem, and the balance-budget theorem. In Metzler (1973a).
1973d. The structure of taxes, open-market operations, and the rate of interest. In Metzler (1973a).
1973e. Partial adjustment and the stability of inventory cycles. In Metzler (1973a).

Bibliography

Gale, D., Nikaido, H. 1965. The Jacobian matrix and the global univalence of mappings. *Mathematische Annalen* 159(2): 81–93.
Haberler, G. 1941. *Prosperity and depression*. 3rd ed. Geneva: League of Nations.
Haberler, G. 1952. The Pigou effect once more. *Journal of Political Economy* 60: 240–246.
Horwich, G. 1962. Real assets and the theory of interest. *Journal of Political Economy* 70: 157–169.
Horwich, G. 1964. *Money, capital, and prices*. Homewood: R.D. Irwin.
Horwich, G., and P.A. Samuelson, eds. 1974. *Trade, stability, and macroeconomics: Essays in honor of Lloyd A. Metzler*. New York: Academic.
McKenzie, L. 1960. Matrices with dominant diagonals and economic theory. In *Mathematical methods in the social sciences 1959*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.
Mundell, R.A. 1960. The public debt, corporate income taxes, and the rate of interest. *Journal of Political Economy* 68: 622–626.
Niehans, J. 1978. Metzler, wealth, and macroeconomics: A review. *Journal of Economic Literature* 16: 84–95.
Patinkin, D. 1956. *Money, interest and prices: An integration of monetary and value theory*. Evanston: Row, Peterson.
Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53: 343–351.
Pigou, A.C. 1947. Economic progress in a stable environment. *Economica* NS 14(55): 180–188.
Scitovszky, T. 1941. Capital accumulation, employment, and price rigidity. *Review of Economic Studies* 8(2): 69–88.
Wright, D.M. 1952. Professor Metzler and the rate of interest. *Journal of Political Economy* 60: 247–249.
Zarnowitz, V. 1985. Recent work on business cycles in historical perspective: A review of theories and evidence. *Journal of Economic Literature* 23: 523–580.

Meynieu, Mary (Died 1877)

A. Courtois

An Englishwoman by birth, married to a Frenchman. She became a widow after 50 years of wedded life, but continued to live in France. Highly cultivated and accomplished, she was one of the few women who have written on political economy, and she was highly successful in popularizing the science.

Her principal works are: *Eléments d'économie politique,* a statement in a series of dialogues between a teacher and pupil for the use of the primary normal schools (Paris, 1839) and *Histoire du paupérisme anglais* (1841).

Microcredit

Jonathan Morduch

Abstract

Most providers of micro-credit lend without requiring collateral. In doing so, they can provide poor households with access to small-scale loans to expand household businesses and meet consumption needs. Micro-credit institutions demonstrate that a combination of mechanisms can overcome the market imperfections created when banks lack good information about borrowers and when borrowers lack collateral. Micro-credit innovations are of both theoretical interest and practical importance. Proponents argue that micro-credit can be a tool to reduce poverty and, in the best cases, can operate profitably and on a large scale, free of public subsidy.

Keywords

Adverse selection; Behavioural economics; Capital access; Collateral; Collusion; Credit

Reprinted from *Palgrave's Dictionary of Political Economy.*

rationing; Grameen bank; Group lending; Information revelation; Joint liability; Micro-credit; Micro-finance; Moral hazard; Muhammad Yunus; Poverty alleviation

JEL Classification
O1

Micro-credit encompasses a broad movement to supply professional banking services to poor households; micro-credit innovations offer new insights into the economics of information and new mechanisms for reducing poverty.

Karl Marx (1867) famously tied inequality in access to capital to broader social and economic inequalities driven by markets; micro-credit presents the promise that market mechanisms may instead help to broaden capital access.

The economics of information shows why poor customers are usually shunned by commercial lenders. Customers typically lack the assets and ownership documents that banks require as collateral, and banks lack cost-effective ways to monitor and enforce contracts. Theorists demonstrate how credit rationing can emerge in these contexts, with adverse selection and moral hazard as culprits. The challenge for banks is exacerbated by the small size of transactions. The *MicroBanking Bulletin's* (2006) survey of 302 leading micro-credit institutions, for example, found that the average loan balance was 436 dollars for the median 'microbank'. For the median micro-bank focusing on poorer customers, the average loan balance was just 109 dollars. These amounts tend to fall below the threshold of interest for large commercial banks, even in low-income economies. Hence the poor lose twice: they begin with less income and fewer assets than others, and, as a result, they have worse access to the financial institutions that might offer a route away from poverty. To this extent, poverty reinforces poverty.

Micro-credit is part of an approach that aims to undo this equation. Despite the challenges, providers of micro-credit aim to deliver reliable and reasonably priced financial services to the under-served, and most institutions aim to do so without ongoing subsidies. While loans are relatively small, advocates argue that the funds are sufficient to finance small businesses and cover emergency consumption needs – and thus to contribute meaningfully to poverty reduction.

Early micro-credit successes were realized in Bangladesh, Indonesia and Bolivia, gaining global attention in the 1980s. By the end of 2005, one annual survey counted over 3000 institutions, collectively serving 113 million customers worldwide (Daley-Harris 2006). Of these, 82 million customers were classified as being among the 'poorest', and 84% of those were women. Rough estimates place unmet demand at over one billion people. The 2006 Nobel Peace Prize to Muhammad Yunus and the Grameen Bank of Bangladesh recognized the potential of micro-credit to reduce global poverty, though Yunus's boldest claims about the potential scale and impact of micro-credit remain untested with reliable data.

The *MicroBanking Bulletin* (2006) survey shows both the promise and the challenges of micro-credit. The survey, which is skewed towards institutions with strong commitments to financial self-sufficiency, finds that 69% of the 302 institutions were earning profits in 2004, and just 2% of loan portfolios were deemed 'at risk' as a result of loan payments left unpaid beyond 30 days. Real interest rates on loans average about 25–35% per year in the survey, though some top 90% per year.

The encouragement of profit and the tolerance of relatively high interest rates are central to the logic of micro-credit policy. Escaping reliance on subsidy, it has been argued, allows institutions to expand beyond the constraints imposed by donors' purses, creating the prospect of a truly global market-based industry. Despite innovations, though, institutions focused on the poorest customers face stiff challenges in generating profits. The *MicroBanking Bulletin* (2006) survey shows that the median micro-bank serving the poorest customers faces almost twice the cost of lending (per unit of assets) compared with the median micro-bank serving betteroff (but still low-income) customers. The extent of trade-offs between meeting profit targets and achieving

social objectives remains largely unexplored, as does the nature of productivity-enhancing roles for subsidy (Armendáriz de Aghion and Morduch 2005, ch. 9).

Group Lending

The high rates of loan repayment are attributed to innovative loan contracts, most notably the 'group lending' contract. The group approach is associated with the Grameen Bank, although it has been employed more faithfully by others. The Grameen approach begins with the bank entering a village and inviting villagers to form themselves into five-person groups. A cluster of groups is then formed into a centre that meets once a week in the village, where all business is transacted by a loan officer from the bank. Loans are given to individuals, but the group is deployed to improve incentives and provide a support network. As long as all loans are repaid on time, loans continue to be made to group members, but if any group member cannot repay his or her loan (and the four others cannot fix the problem themselves), the entire group is excluded from future borrowing. This element of the contract is often referred to as creating 'joint liability' – even though, in the Grameen model at least, individuals are not explicitly liable for the repayment of fellow group members.

The contract addresses moral hazard by giving borrowers incentives to monitor each other and to sanction members whose lack of effort jeopardizes loan repayments. The customers often have advantages in these activities, which stem from living and working alongside each other and from being able to employ 'social' sanctions that the bank cannot use. The contracts may also foster mutual support mechanisms that provide insurance and other assistance, a point stressed by Muhammad Yunus, Grameen's founder. Early theoretical analyses on moral hazard in group lending include Stiglitz (1990), while Besley and Coate (1995) raise the possibility of collusive behaviour by borrowers.

The contracts, in principle, can also address adverse selection (and the inefficiencies created by the withdrawal of safer borrowers in markets with asymmetric information; Stiglitz and Weiss 1981). Adverse selection in credit markets arises because banks cannot distinguish between potential customers who are likely to reliably repay loans and those that will not. Without such information, the bank must charge all customers the same interest rate, and the safer borrowers implicitly subsidize the riskier ones.

In principle, the process of group formation can improve outcomes by screening risky borrowers and matching safer individuals with other relatively safe individuals. Because the effective cost of the loan depends in part on the probability that one's fellow group members will default, safer individuals will then face lower effective borrowing costs than riskier individuals – even when all individuals face identical nominal contracts; the contract combined with the sorting process reduces the extent of cross-subsidization and thus adverse selection (Ghatak 1999; see also references in Armendáriz de Aghion and Morduch 2005, ch. 3). This mechanism has received little empirical verification, though, and in practice lenders devote substantial resources to information acquisition.

Beyond Group Lending

The use of groups has clear attractions but has proved cumbersome when customers have diverse needs and growth prospects. It also relies on the willingness and ability of customers to carry out monitoring and enforcement tasks that are usually the responsibility of bankers. Rai and Sjöström (2004) point to inefficiencies in group lending that can be mitigated through simple information revelation mechanisms, and, as noted above, collusion remains a theoretical possibility. A push to move beyond group lending with joint liability reached an important milestone in practice when two early pioneers, BancoSol of Bolivia and the Grameen Bank, independently abandoned group lending with joint liability as the basis of their operations.

The move beyond group lending highlights other contract innovations that have been overshadowed. Among the most important is the repayment schedule. In Grameen Bank loan

contracts, for example, loans are repaid in small increments weekly over the course of several months to a year. It is an odd structure for loans that are ostensibly for business investments that may take time to bear fruit. The schedule, though, allows households to repay loans from other income sources in small, manageable increments. In this way, loans can often be repaid even if businesses fail. Perhaps more important, the structure allows households to easily use loans to finance consumption, strengthening the ability to cope with health crises, pay school fees and keep food on the table. The extent to which such 'diversion' occurs, and its costs and benefits, has yet to receive much research attention, but it may hold a key to new directions for micro-credit.

A second important mechanism is the use of long-term lending relationships. Lenders gain information and instill incentives for loan repayment by repeatedly interacting with customers, allowing borrowers to start with small loans and become eligible for steadily larger loans with each successful cycle.

From Micro-Credit to 'Micro-Finance'

Most of the evidence in favour of micro-credit is anecdotal, though rigorous empirical studies are accumulating (Armendáriz de Aghion and Morduch 2005, ch. 8). In data from Mexico, for example, McKenzie and Woodruff (2006) find returns to capital of above 20% per month for small-scale businesses with capital stocks below 200 dollars. As capital stocks rise above 400 dollars, estimated returns to capital fall to around 5% per month. These returns are still substantial and help to explain the ability to pay relatively high interest rates.

The returns pose a puzzle, though. There are no signs of poverty traps, and if returns are so high, why have households been unable to save more on their own, overcoming credit gaps through self-finance? With the realization that customers indeed seek better ways to save and insure (and seek credit for a wide range of uses), micro-banks have started expanding their services. The next wave of innovations focuses there, and draws in part on insights from behavioural economics (for example, Ashraf et al. 2006). The focus will thus continue to shift from 'micro-credit' to 'micro-finance' more broadly.

See Also

- ▶ Adverse Selection
- ▶ Credit Rationing
- ▶ Development Economics
- ▶ Moral Hazard
- ▶ Poverty Alleviation Programmes
- ▶ Poverty Traps

Bibliography

Armendáriz de Aghion, B., and J. Morduch. 2005. *The economics of microfinance*. Cambridge, MA: MIT Press.

Ashraf, N., D. Karlan, and W. Yin. 2006. Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics* 121: 635–672.

Beck, T., A. Demirguc-Kunt, and M.S. Martinez Peria. 2006. *Banking services for everyone? Barriers to bank access and use around the world*, World bank policy research working paper 4079. Washington, DC: World Bank.

Besley, T., and S. Coate. 1995. Group lending, repayment incentives, and social collateral. *Journal of Development Economics* 46: 1–18.

Daley-Harris, S. 2006. *State of the microcredit summit campaign report 2006*. Washington, DC: Microcredit Summit.

Ghatak, M. 1999. Group lending, local information and peer selection. *Journal of Development Economics* 60: 27–50.

Marx, K. 1867. *Capital*, vol. 1, tr. Ben Fowkes. Harmondsworth: Penguin, 1990

McKenzie, D., and C. Woodruff. 2006. Do entry costs provide an empirical basis for poverty traps? Evidence from Mexican microenterprises. *Economic Development and Cultural Change* 55: 3–42.

Microbanking Bulletin 12. 2006. Washington, DC: Microfinance Information Exchange

Rai, A., and T. Sjöström. 2004. Is Grameen lending effcient? Repayment incentives and insurance in village economies. *Review of Economic Studies* 71: 217–234.

Stiglitz, J. 1990. Peer monitoring and credit markets. *World Bank Economic Review* 4: 351–366.

Stiglitz, J., and A. Weiss. 1981. Credit markets with imperfect information. *American Economic Review* 71: 393–410.

Microeconomics

Hal R. Varian

Microeconomics is the study of individual economic units and their interactions. In includes the theory of the consumer, the producer, and the markets in which they are involved. Microeconomics is often contrasted with macroeconomics which is concerned with the behaviour of economic aggregates, such as aggregate consumption and production.

These days the distinction between macroeconomics and microeconomics is becoming rather blurred. Considerable work in recent years has gone into investigating the 'microeconomic foundations of macroeconomics' and much current research in macroeconomics has a distinctly micro flavour. Still, the goal of macroeconomics seems to be to understand and predict the behaviour of aggregate economic variables – consumption, investment, employment, etc. – rather than understand a single economic unit or market in isolation, so this difference in focus may serve as a distinguishing characteristic.

An analogy with thermodynamics is instructive. There we have the 'micro-theory' of statistical thermodynamics which begins by studying the behaviour of individual molecules and their interactions, and then derives the implications of this behaviour for the system as a whole. This is to be contrasted with the 'macro-theory' of classical thermodynamics, which postulated three laws of thermodynamic systems, and derived their implications. Both theories attempt to describe the same behaviour, albeit from a different approach.

Similarly, in microeconomics the starting point is the individual decision-maker and his choices. The aggregate relationships are meaningful only as the sum of the individual decisions. In macroeconomics, on the other hand, one often begins by postulating functional relationships among aggregate economic quantities (consumption, investment, money,. . .) and then proceeds to examine their interactions.

Origin and Use of the Term

The origins of the terms microeconomics and macroeconomics are surprisingly obscure. Machlup (1963) says that 'Ragnar Frisch used [the terms] as far back as 1933 and it was probably he who introduced them into economics.' But the work of Frisch's that Machlup cites does not contain either word; instead, Frisch used the words 'micro-dynamic' and 'macro-dynamic', albeit in a way closely related to the current usage of the terms 'microeconomic' and 'macroeconomic':

> The micro-dynamic analysis is an analysis by which we try to explain in some detail the behaviour of a certain section of the huge economic mechanism, taking for granted that certain general parameters are given . . . The macrodynamic analysis, on the other hand, tries to give an account of the whole economic system taken in its entirety (Frisch 1933).

Elsewhere Frisch gives a more explicit definition of these terms that is closely akin to the modern usage of micro and macroeconomics: 'Micro-dynamics is concerned with particular markets, enterprises, etc., while macro-dynamics relate to the economic system as a whole' Frisch (1934).

The distinction between macro- and microeconomics became quite popular after the publications of Keynes's *General Theory* in 1936. Although Keynes does not seem to have used either of the terms himself, he was acutely aware of the distinction, as indicated in the following passage:

> The division of Economics between the Theory of Value and Distribution on the one hand and the Theory of Money on the other hand is, I think, a false division. The right dichotomy is, I suggest, between the Theory of the Individual Industry or Firm and of the rewards and the distribution of a *given* quantity of resources on the one hand, and the Theory of Output and Employment *as a whole* on the other hand (Keynes 1936, p. 293; emphasis in original).

The earliest published reference that explicitly uses the term 'microeconomics' that I have been able to locate is de Wolff (1941). De Wolff, a colleague of Tinbergen at the Netherlands Statistical Institute, was well aware of the macro-dynamic modelling efforts of Frisch, and may have been inspired to extend Frisch's use of 'micro-dynamics' to the more general expression

of 'microeconomics'. De Wolff's note is concerned with what we now call the 'aggregation problem' – how to move from the theory of the individual consuming unit to the behaviour of aggregate consumption. De Wolff correctly points out that only in very special conditions will this sort of aggregation be independent of the distribution of income, and calls for a more careful study of the relationship between micro and macro specifications of economic behaviour. He is quite clear about the distinction between micro- and macroeconomics:

> The concept of income elasticity of demand has been used with two entirely different meanings: a micro- and macro-economic one.
>
> The micro-economic interpretation refers to the relation between income and outlay on a certain commodity *for a single person or family.*
>
> The macro-economic interpretation is derived from the corresponding relation between total income and total outlay *for a large group of persons or families* (social strata, nations, etc.) (de Wolff 1941, p. 140; emphasis in original).

Thus it appears that the use of the term microeconomics probably evolved from the work of Frisch and Tinbergen in the mid-Thirties; and that de Wolff was one of the earliest authors to describe the distinction between micro- and macroeconomics in print.

Despite this mystery surrounding the coinage of the words, by the mid-Forties the terms were beginning to appear to academic journals (e.g., Klein 1946) and textbooks (e.g. Boulding 1948). However, the distinction did not make its way into Samuelson's famous introductory text until 1958. Since the mid-Fifties, the terms have been in widespread use.

Other Interpretations of the Distinction

The distinction between microeconomics and macroeconomics described above is the standard one, but there are more subtle issues involved. This is most clearly indicated in the quote from Keynes given earlier. It is apparent here that Keynes is not only concerned with the 'individual' vs. 'aggregate' distinction, but also with the 'full employment' vs. 'underemployment' distinction.

Some authors have argued that the fundamental distinction between macroeconomics and microeconomics is (or should be?) that macroeconomics is an attempt to understand situations of underemployment and excess capacity, while microeconomics is primarily concerned with situations of full utilization of resources. For an interesting discussion of this viewpoint, see Leijonhufvud (1968).

This distinction is well taken; certainly the inspiration for what we today call macroeconomics came from an attempt to understand the prolonged underutilization of resources during the 1930s. However, it is no contradiction in terms to speak of 'full employment macroeconomics' or 'microeconomic theories of unemployment'. Most economists would presume that the first phrase would deal with equilibrium models expressed in aggregate terms, while the second would examine reasons for the existence of unemployment based on the behaviour of individual workers and firms. Thus it seems that the aggregated vs. disaggregated nature of the study is the fundamental distinction in the ordinary use of the terms.

The Methods of Microeconomics

Economics proceeds by building models of behaviour. These models are supposed to be simplified representations of reality which specify how variables in a system relate to each other. Economists use many techniques in the construction and analysis of economic models, but most of the techniques fall into the categories of optimization analysis and equilibrium analysis.

Nearly all models of individual behaviour in microeconomics are models of optimizing behaviour. Indeed, some economists have gone so far as virtually to identify optimizing behaviour with 'rational behaviour'. In building a model of behaviour, economists are naturally led to identify agents that make choices, the kinds of choices that are feasible for them, how the choices of other agents constrain them, and so on. Once the economist is able to write down an optimization problem describing an economic choice, he or she can

bring the powerful mathematical methods of microeconomic analysis to bear.

These mathematical tools allow the economist to use standard results to discuss conditions under which the optimization problem has a solution, when it is unique, and how it varies with the underlying parameters. The hypothesis that the behaviour under examination is optimizing behaviour allows one to draw nontrivial inferences about how choices will respond to changes in the economic environment. This type of exercise is known as 'comparative statics', although a better name for it would probably be 'sensitivity analysis'.

Once we have understood the nature of the optimal choice problem facing individual agents, we can investigate how these choices fit together. In general, some of the variables that influence a given agent's behaviour – such as prices – will be determined, at least in part, by the behaviour of other agents. An economic equilibrium is a situation where no agent has an incentive to change any of his choices, given his perceptions of the behaviour of the other agents.

Since the optimal choice of one agent can be expressed as a function of the optimal choice of the other agents, the study of economic equilibrium will generally reduce to the study of the solution of a simultaneous set of equations. The fact that these equations are solutions to optimization problems will impose some structure on the system, and the modelling choices of which variables are involved and how they enter the choice problem will add additional restrictions. Again, there are standard mathematical techniques that can be used to examine the existence, uniqueness and comparative statics properties of such systems of equations.

The paradigm example of this sort of microeconomic analysis is the study of competitive equilibrium. First we examine the individual choices of households and firms. Each household and firm is assumed to take as given the prices of the various goods that it consumes and produces. Households are assumed to maximize utility, and firms are assumed to maximize profits. Once these maximization problems have been posed, the economist can analyse the choice behaviour of consumers and firms using the mathematical techniques alluded to above.

We can then proceed to specify the equilibrium conditions of the model. In the case of competitive equilibrium, it is postulated that the price of a good will not change when the total demand for that good equals supply. One is then led to seek conditions under which such an equilibrium exists, and to understand how it changes as underlying parameters of the households and firms change.

As described above, the building of a microeconomic model may seem quite mechanical. Nevertheless there is an art to it. It is easy to write down a model in which everything depends on everything else, and only trivial conclusions emerge. The art in model building consists in knowing what to leave out. As mentioned above, a model is a simplified representation of reality, and the decision of what simplifications to make is the essence of building an economic model. A map on a one-to-one scale is useless, as is an economic model that attempts to describe every aspect of economic reality. Only by eliminating irrelevant detail can we hope to gain an understanding of complex social and economic processes.

The Current Status and the Future of Microeconomics

The methods described above have been applied with great success to the classical theory of the firm and the consumer. It is now safe to say that virtually all of the useful consequences of maximizing behaviour in these contexts have been well understood. In any intellectual field of inquiry there are subjects that are mature and subjects that are in their infancy, and the classical theory of the consumer and the firm must be counted as mature subjects.

Similar progress has been made in examining the behaviour of competitive markets, and to a lesser extent, models of imperfect competition. In the last decade it has become apparent that successful analysis of imperfect competition requires a treatment which explicitly uses the

methods of game theory, and such investigations are currently flourishing.

Recently, there has been great interest in the microeconomics of information, and there have been many important advances in this area. Conventional models typically assume full information on the part of all economic agents. When we relax this assumption and allow agents to have different information, new and surprising properties emerge. In these models one wants an equilibrium not only in the decisions by the agents to acquire goods – for example their demands and supplies – but also in their decisions to acquire information. These extra equilibrium conditions can be quite important in determining the behaviour of the economic system.

With so much energy being applied to the economics of imperfect information and game theoretic models of imperfect competition, I expect that many of the problems in this area that now seem so formidable will be solved in the not too distant future (or be shown to be intractable!). These subjects are today very lively areas of research; if the theory of the consumer and the producer are mature subjects, we might say that the study of imperfect competition and imperfect information are in their adolescence.

What does the future hold? It seems to me that the current 'infant' subject of microeconomics is the research that is examining the 'micro-microeconomics' of firms, consumers and markets. By this I mean the investigations that attempt to go behind the 'black box' of the neoclassical firm, consumer and market, and try to understand the internal functioning of these economic institutions.

In terms of the consumer, there has been considerable interest in the internal structure of the household as a consuming unit, and, more recently, economists have been probing into the internal structure of individuals using psychologically based models of choice behaviour.

In terms of the firm, there is much promise in the recent work about the internal organization of production. The new view of the firm as a 'nexus of contracts' will, when suitably formalized and digested, undoubtedly lead to a more profound understanding of real world firm behaviour.

Finally, the last few years have seen some significant progress in understanding the internal functioning of markets. Economic theorists are attempting to model rational price setting agents in competitive, or nearly competitive markets, and to come to grips with key issues of microeconomic dynamics.

We began this essay by defining microeconomics as the study of individual economic units such as consumers, firms and markets, and how they interact. If this is so, then the areas described immediately above might be thought of as 'nanoeconomics' – the study of the constituent parts of consumers, firms and markets, and how these units interact in the larger economic environment. I suspect that progress in microeconomic *and* macroeconomic theory in the future will rest, to a large extent, on a deeper understanding of such phenomena.

See Also

► Macroeconomics: Relations with Microeconomics

Bibliography

Boulding K. 1948. *Economic analysis.* Revised ed. New York: Harper and Brothers.

De Wolff, P. 1941. Income elasticity of demand, a microeconomic and a macroeconomic interpretation. *Economic Journal* 51: 140–145.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honour of Gustav Cassel*, ed. R. Frisch. London: Allen & Unwin.

Frisch, R. 1934. Some problems in economic macrodynamics. *Econometrica* 2: 189.

Keynes, J.M. 1936. *The general theory of employment, interest and money.* New York: Harcourt, Brace.

Klein, L. 1946. Macroeconomics and the theory of rational behavior. *Econometrica* 14: 93–108.

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. New York: Oxford University Press.

Machlup, F. 1963. Micro- and macro-economics: Contested boundaries and claims of superiority. In *Essays on economic semantics*, ed. F. Machlup and M. Miller. Englewood Cliffs: Prentice-Hall.

Microfoundations

Maarten C. W. Janssen

Abstract

The microfoundations literature has attempted to bridge the gap between microeconomic and macroeconomic models. Many models in this literature have used the theoretical construct of a representative agent. Economy-wide outcomes are thereby presented as if they were the result of the optimizing behaviour of one individual. Emergent properties at the macro level are by construction precluded from the analysis. Other literatures exist where emergent properties are taken to be at the heart of the quest for microfoundations.

Keywords

Aggregation (theory); Austrian economics; Behavioural macroeconomics; Bounded rationality; Classical economics; Coordination failure; Effective demand; Efficiency wages; Evolution; Expectations; Fairness; Fixprice models; Frictions; Game theory; General equilibrium; Hayek, F. von; Imperfect competition; Involuntary unemployment; IS–LM; Keynesianism; Labour supply; Learning; Lucas, R; Market clearing; Menu cost; Methodological individualism; Microfoundations; Multiplier; Neoclassical synthesis; New Classical economics; New Keynesian macroeconomics; Other-regarding preferences; Overlapping generations model of general equilibrium; Phillips curve; Post-Keynesian economics; Residential segregation; Rational behaviour; Rational expectations; Real business cycles; Reciprocity; Representative agents; Spontaneous order; Sticky prices; Strategic complementarities; Technology shocks; Vienna Circle

JEL Classification

D0; E0

The quest to understand microfoundations is an effort to understand aggregate economic phenomena in terms of the behaviour of individual economic entities and their interactions. These interactions can involve both market and non-market interactions. The quest for microfoundations grew out of the widely felt, but rarely explicitly stated, desire to stick to the position of methodological individualism (see Agassi 1960, 1975; Brodbeck 1958), and also out of the growing uneasiness among economists in the late 1950s and 1960s with the coexistence of two sub-disciplines – namely, microeconomics and macroeconomics – both aiming to explain features of the economy as a whole. Methodological individualism is the view that proper explanations in the social sciences are those that are grounded in individual motivations and their behaviour. The urge to make microeconomics and macroeconomics compatible can be understood from the perspective of the unity-of-science discussion initiated by the Vienna Circle in the philosophy of science in the beginning of the twentieth century (see Nelson 1984).

Efforts to understand microfoundations go far beyond the questions that lie at the heart of formal aggregation theory, that is, the analysis of how to map aggregate economic variables and relationships back to similar individual variables and relationships that underlie them. One crucial issue in the microfoundations literature is the extent to which aggregate economic variables and/or relationships exhibit features that are similar to the features of individual variables and/or relationships, and in particular whether certain features are emergent properties at the macro level that do not have a natural counterpart at the individual level. An important early example of emergence is Schelling's analysis (1978) of segregation. He shows that segregation in neighbourhoods may be an emergent property at the micro level that can be viewed as an unintended consequence of the individual decisions concerning where to live.

The discussion on emergence shows that there is no reason to assume or expect macro behaviour to be in any way similar or analogous to the behaviour of individual units. In order to have 'proper' microfoundations in line with

methodological individualism, it is thus by no means required that aggregate outcomes are represented as if they were the outcome of a single agent's decision problem. On the contrary, the restriction to single individual decision problems found in modern macroeconomics is self-imposed and not implied by the methodological position of methodological individualism (see Kirman 1989). In fact, one may argue that the interaction between different, and possibly heterogeneous, individual units should be at the core of macroeconomic analysis.

As the quest for 'proper' microfoundations has arisen in the debate concerning the microfoundations for macroeconomics, this article's main focus is on this debate. The article starts with a historical perspective on this debate and continues to discuss New Classical and New Keynesian approaches to macroeconomics that emerged out of the microfoundations debate. The role of equilibrium notions and expectations is discussed in a separate section. The article argues that the microfoundations for macroeconomics literature is best understood from the perspective of attempting to make microeconomics and macroeconomics compatible with each other. The article closes with a discussion of non-mainstream approaches to microfoundations and more recent approaches to microfoundations using the perspective of evolutionary forces and boundedly rational behaviour.

Historical Background to the Microfoundations for Macroeconomics Debate

Around the mid-1950s two more or less separate approaches existed to studying economy-wide phenomena: general equilibrium theory and (Keynesian) macroeconomics. Some of the more important theoretical issues within each of these approaches were settled. Existence of a general equilibrium point was proved by Arrow and Debreu (1954) and the macroeconomic IS–LM framework was well established (following the seminal paper by Hicks 1937). Of course, some other issues were still to be tackled, such as questions related to how to deal with imperfect competition, incomplete markets and/or overlapping generations.

Both approaches explained economy-wide phenomena, but there were important differences between the perspectives from which they started. Flexible prices and market-clearing were at the core of general equilibrium theory; involuntary unemployment and effective demand were important concepts in macroeconomics. The neoclassical synthesis reconciled general equilibrium theory and (Keynesian) macroeconomics by giving each of them its own domain of applicability: macroeconomics (with its assumption of sticky money wages) gives an accurate description of the economy in the short run, while long-run developments of the economy were considered to be adequately described by the general equilibrium approach.

From a theoretical point of view this state of affairs was unsatisfactory. One cannot simply attribute unemployment to sticky money wages while leaving the theoretical structure of general equilibrium theory intact: the imposition of a fixed money wage (or, more generally, fixed prices) deeply affects the theory of supply and demand. It was natural, then, to inquire into the relationship between the two approaches, especially given that they study the same phenomena. In addition, the generally accepted view was that it is the market interaction between many individual agents from which economy-wide phenomena result, implying that general equilibrium theory is the more fundamental theory of the two. The quest for microfoundations was born.

The rise of interest in microfoundations can also be at least party conceived as being driven by the perceived failings of important elements of empirical macroeconomics and in particular the fact that the Phillips curve turned out to be not a stable relationship that can be used for economic policy purposes (see, for example, Friedman 1968). Several essays in Phelps (1970) are written to reconcile microeconomic theory with the apparent temporary trade-off between wages and unemployment embodied in the new interpretation of the Phillips curve.

New Classical and New Keynesian Economics

One key controversy in the quest for microfoundations is how to explain the widely observed phenomenon of unemployment. From a market-clearing perspective, unemployment simply means that at the current (real) wage rate people do not want to supply more labour to the market. If there is registered unemployment, it is thus either of a 'voluntary' nature or a short-run phenomenon that quickly disappears. In this vein, Lucas (1978, p. 354) argued that involuntary unemployment is not a fact that needs to be explained, but rather a theoretical construct Keynes introduced in the hope it would be helpful in explaining fluctuations in measured unemployment.

In line with these ideas, New Classical economists have attempted to reconcile macroeconomic phenomena such as inflation and unemployment, and the empirical observed trade-off between the two measured by the Phillips curve, with a Walrasian notion of market clearing. Early models, such as Lucas and Rapping (1969) and Lucas (1972), stressed the idea that incomplete information about the money supply may cause business fluctuations. Later real business cycle models (such as that of Kydland and Prescott 1982) looked at technology shocks to explain cyclical behaviour. Thus, an important difference between the Lucas–Rapping approach and early real business cycle models is that the former, but not the latter, introduces frictions to explain business cycles. With these New Classical models, the concept of the representative agent (consumer, firm or producer/consumer agent) became widely used in modern macroeconomics. In its most extreme form, the economy as a whole is represented as if it were the outcome of a single individual's decision problem. The possible differences between individual and aggregate economic behaviour are thereby assumed away.

Economists who were oriented towards Keynesian ideas thought that there is an involuntary, non-transient component in observed unemployment figures. Many New Keynesian contributions therefore try to reconcile the notion of involuntary unemployment with a notion of market equilibrium.

A first approach considers the question how to incorporate the notion of price stickiness, especially concerning money wages, with the traditional theory of demand and supply. This issue was first studied by Clower (1965). He emphasized that, because of the interdependence of markets, demand and supply curves on all markets are affected if money wages are fixed. If prices are restrained from bringing about market clearing allocations, then other variables have to bring about some kind of fixed-price equilibrium. Clower (1965) and Leijonhufvud (1968) set out a research programme studying the existence of fixed-price equilibria and their properties. The resulting equilibrium notion and the properties of such fixprice equilibria were formulated by Barro and Grossman (1971), Drèze (1975) and Benassy (1975), among others. The idea of this literature is that agents express their demands on the basis of market prices and perceived quantity constraints. These models have microfoundations in the sense that they are based on decision-making individuals and a notion of equilibrium. Moreover, it turned out that the fixprice models capture quite a number of ideas associated with Keynesian economics. By means of these alternative equilibrium notions, involuntary unemployment could be regarded as an equilibrium phenomenon in which optimizing households face a quantity constraint on the amount of labour they can supply. Also, the Keynesian notions of effective demand and the multiplier were reformulated within the new models. Finally, the models provided arguments for demand policies by the government. Of course, from a market-clearing perspective, these fixprice models are unsatisfactory as they do not explain why (rational) individuals do not propose changes to the terms of trade at which they exchange. Clearly, if prices are fixed at no market clearing levels, some agents in the economy can mutually benefit by exchanging at different prices, and therefore have an incentive to propose changes in prices. A literature on small menu cost appeared arguing that introducing a very small cost for economic agents to change prices

may result in large fluctuations in aggregate output (see Mankiw 1985).

Another approach New Keynesian economists followed is to incorporate the literature on imperfect competition into macroeconomic models. Hart (1982), Blanchard and Kiyotaki (1987), Kiyotaki (1988) and d'Aspremont et al. (1990) are among the pioneering articles in this area. These models can explain why aggregate output is below the optimal full employment output level. Unemployment can be involuntary when there is imperfect competition in the labour market.

A third approach to explaining non-competitive wages is to introduce some type of informational problem, as in the literature on efficiency wages. The basic idea of this literature is that the average labour productivity is positively related to the wage a firm offers. Firms may set wages above the competitive level in order to induce employees to work harder, and therefore may be unwilling to lower their wage offers (see Yellen 1984; Lindbeck and Snower 1987).

Yet another approach relies on coordination failures formally analysed in terms of multiple equilibria (see Bryant 1983; Roberts 1987). Cooper and John (1988) point out that many New Keynesian models are based on strategic complementarities between agents' actions, that is, these models do not rely on an assumption that prices cannot adjust to their market equilibrium values. When strategic complementarity exists, there may be multiple equilibria that can be Pareto-ranked. Agents may then find themselves in a 'bad' equilibrium, but individually they cannot benefit by deviating to another choice. The authors call this a 'coordination failure'.

There is a parallel between the coordination failures literature and the overlapping generations general equilibrium literature (see, for example, Geanakoplos and Polemarchakis 1986). The latter literature views the economy as a process without definite end, such that what happens today is underdetermined as it depends on what people expect to happen tomorrow, which in turn depends on what people expect to happen the day after tomorrow, and so on. In such a world there is a continuum of equilibria. Geanakoplos and Polemarchakis (1986) show that, depending on how this indeterminacy is solved, that is, which variables are chosen to be exogenously determined, classical or Keynesian-oriented conclusions may be derived.

Work on all these different models has resulted in a shared methodology of how to go about building macroeconomic models. The traditional distinction in macroeconomics between Keynesian and classical economists is disappearing and a common methodology is surfacing. Economists share the understanding that the ultimate question that matters is how well markets function. The differences in importance attached to various market frictions are more a matter of degree than of fundamental divergence between different methodologies. The nature of what used to be macroeconomic theory has undergone dramatic changes alongside these developments. Traditional macroeconomic issues such as how to explain the business cycle or how to account for inflation are now studied with the same tools and techniques as those that are used in microeconomics. Along these lines, and by using the assumption of the representative agent, modern macroeconomics has assumed away the heterogeneity that may exist at the individual level. Lucas's prediction that we may soon simply speak of economic theory instead of separate microeconomic and macroeconomic theories has turned out to be fairly accurate (see Lucas 1987, pp. 107–8). Somewhat paradoxically, one may say that the modern economist who still is a 'hard line microeconomist' is now called a macroeconomist.

Rationality, Equilibrium and Expectations

The efforts to create microfoundations for macroeconomics have resulted in a more unified approach to doing economic theory. The approaches discussed so far (also Keynesian-oriented models) all postulate rational behaviour on the part of economic agents and some notion of equilibrium. If expectations are important, it is postulated that agents' expectations concerning

important variables coincide with the model's predicted values concerning these same variables. This assumption concerning agents' expectations have been termed 'rational expectations' (see Muth 1961).

Parallel to the microfoundations literature, a literature questioning the eductive justifications for the notions of equilibrium and rational expectations emerged. This literature on the foundations of game theory basically argued that, if we assume that agents (players) are rational and that their rationality and the model (game) in which they operate are common knowledge, then it is not implied that these agents will play according to an equilibrium of the game. Fundamental papers in this respect are Bernheim (1984) and Pearce (1984), among others. These and other papers show that a much weaker notion, named (correlated) rationalizability, can be derived from assumptions regarding common knowledge of the rationality of players.

On the basis of this literature, Guesnerie (1992) argues that rational expectations should be regarded as an equilibrium notion that is also not solely based on postulates regarding the rational behaviour of individual players. It is rational for individual players to have 'rational expectations' if other players have these very same 'rational expectations', but not necessarily otherwise. As the notion of rational expectations is essentially an equilibrium or consistency notion, it suffers from the same drawbacks in that it is not implied by the individual rationality assumptions that players will form rational expectations.

Another literature (see, for example, Bray and Savin 1986, and several essays in Frydman and Phelps 1983) studies the question whether in a decentralized economy economic agents may learn over time to have expectations that are consistent with those that are assumed by the rational expectations hypothesis. The general conclusion of this literature is that, due to the feedback from expectations to economic behaviour, the outcomes of an economic model with learning agents do not converge to the rational expectations solution.

It then follows that the microfoundations literature mentioned so far has not really succeeded in deriving all macroeconomic propositions from fundamental hypotheses on the behaviour of individual agents. The requirements of methodological individualism have thus not been satisfied by the microfoundations literature, which has predominantly presumed that individuals behave rationally (see Janssen 1993).

Non-mainstream Approaches to the Microfoundations of Macroeconomics

Apart from a long-lasting debate in the mainstream literature, the term 'microfoundations' has also stimulated work by other economists, and they have publicized their views on the relation between microeconomics and macroeconomics. Horwitz (2000) provides an overview of the Austrian perspective where individual knowledge, prices as conveyers of information, and subjective evaluations play important roles. The essays in Hayek (1948) and his views on spontaneous order are especially important in this respect. It may seem, then, that macroeconomics is not an important term in the Austrian vocabulary. However, this is only partly true. From an Austrian perspective an important question is what kind of monetary system will most likely preserve the communicative function of prices. Austrian economists have, as Horwitz shows, addressed such issues in a way that is compatible with methodological individualism.

A post-Keynesian view of the economy holds that long-term expectations are largely determined by non-economic processes such as those determined by mass psychology. These expectations therefore should be regarded as exogenous to the economic model, rather than as endogenously determined as in the case of rational expectations. Interestingly, this post-Keynesian view comes close to the result that is established by Geanakoplos and Polemarchakis (1986) in their overlapping generations general equilibrium model, where they show that indeterminacy of

equilibria implies that expectations concerning future market outcomes may be chosen exogenously. Important investment decisions are, according to post-Keynesian economists, by their nature long-term decisions, and these decisions are thus largely determined by the state of these long-term expectations. This fundamental uncertainty requires a different decision-theoretic approach from what is typically used by mainstream economics. Informally, some post-Keynesians have argued for the irreducibility of macroeconomic issues to purely microeconomic considerations where individuals' actions are based on expected utility calculations (see Weintraub 1979).

Alternative Types of Microfoundations

Most of the literature up to the 1990s discussing the microfoundations of macroeconomics has focused on rationally behaving self-interested economic agents. More recently, attention has shifted to other forms of behaviour. Using evolutionary mechanisms or learning, economists have studied the evolutionary foundations of equilibrium notions (see Kandori et al. 1993; Young 1993). Allowing agents to imitate best practices they observe around them, or choosing best replies to some adaptively formed expectations of what others will do, the literature shows that under some conditions concerning the dynamic process the economy will converge to equilibrium play. Early work in this direction by Schelling (1978) shows, as noted in the introduction to this article, that macro phenomena such as racial segregation may be regarded as the unintended long-run outcome of the interactive effects of decisions of individual households to move into other neighbourhoods.

Alternatively, economists such as Fehr and Falk (1999) have looked at the consequences of non-selfish preferences for macroeconomic outcomes. They consider preferences for fairness and reciprocity to be important in explaining why managers do not consider cutting employees' wages. Wage cuts may be perceived as unfair and hostile, and managers fear that they will be followed by hostile actions on the part of employees. This literature provides an alternative foundation for the downward rigidity of monetary wages, and may start a literature on behavioural macroeconomics.

Conclusions

The microfoundations literature has brought about many changes in economic theory. Macroeconomic theory in the form of studies of the interplay of a few aggregate relationships is almost non-existent nowadays. Instead, an extreme form of 'microfoundations' is sometimes used in which the economy as a whole is represented in terms of a single agent decision problem. In this way, emergent properties appearing at the macro level that do not exist at the individual level are precluded from the analysis as the micro and macro level simply coincide!

Along with the many other models in the microfoundations literature reviewed in this article, we now see a wide spectrum of partly overlapping models dealing with different types of market frictions and market imperfections. Most of the literature before the 1990s adopts fairly traditional assumptions concerning individual behaviour. More recent contributions in the area of behavioural economics and evolutionary models with (adaptively) learning individuals are starting to explore the implications of different behavioural assumptions at the individual level and to consider the macro implications. These models have the potential to analyse how macro phenomena may emerge from the interactions among a heterogeneous set of individuals. Thereby, they may provide economic theory with a more plausible empirical underpinning, while sticking to the requirements of methodological individualism.

See Also

- ▶ Involuntary Unemployment
- ▶ Methodological Individualism
- ▶ Social Interactions (Theory)

Bibliography

Agassi, J. 1960. Methodological individualism. *British Journal of Sociology* 11: 144–170.

Agassi, J. 1975. Institutional individualism. *British Journal of Sociology* 26: 144–155.

Arrow, K., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Barro, R., and H. Grossman. 1971. A general disequilibrium model of income and employment. *American Economic Review* 61: 82–93.

Benassy, J.-P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42: 502–523.

Bernheim, D. 1984. Rationalizable strategic behavior. *Econometrica* 52: 1007–1028.

Blanchard, O., and N. Kiyotaki. 1987. Monopolistic competition and the effects on aggregate demand. *American Economic Review* 77: 647–666.

Bray, M., and N. Savin. 1986. Rational expectations equilibria, learning and model specification. *Econometrica* 54: 1129–1160.

Brodbeck, M. 1958. Methodological individualisms: Definition and reduction. *Philosophy of Science* 25: 1–22.

Bryant, J. 1983. A rational expectations Keynes type model. *Quarterly Journal of Economics* 98: 525–529.

Clower, R. 1965. The Keynesian counterrevolution: A theoretical appraisal. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Cooper, R., and A. John. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103: 441–463.

d'Aspremont, C., R. Dos Santos Ferreira, and L. Gérard-Varet. 1990. On monopolistic competition and involuntary unemployment. *Quarterly Journal of Economics* 105: 895–919.

Drèze, J. 1975. Existence of an equilibrium with price rigidity and quantity rationing. *International Economic Review* 16: 301–320.

Fehr, E., and A. Falk. 1999. Wage rigidity in a competitive incomplete contract labour market. *Journal of Political Economy* 107: 106–134.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Frydman, R., and E. Phelps (eds.). 1983. *Individual forecasting and aggregate outcomes*. Cambridge: Cambridge University Press.

Geanakoplos, J., and H. Polemarchakis. 1986. Walrasian indeterminacy and Keynesian macroeconomics. *Review of Economic Studies* 53: 755–779.

Guesnerie, R. 1992. An exploration of the eductive justifications of the rational expectations hypothesis. *American Economic Review* 82: 1254–1278.

Hart, O. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97: 109–138.

Hayek, F. 1948. *Individualism and economic order*. Chicago: Chicago University Press.

Hicks, J. 1937. Mr. Keynes and the classics: A suggested interpretation. *Econometrica* 5: 147–159.

Horwitz, S. 2000. *Microfoundations and macroeconomics: An Austrian perspective*. London: Routledge.

Janssen, M. 1993. *Microfoundations: A critical inquiry*. London: Routledge.

Kandori, M., G. Mailath, and R. Rob. 1993. Learning, mutation and long-run equilibria in games. *Econometrica* 61: 29–56.

Kirman, A. 1989. The intrinsic limits of modern economic theory: The emperor has no clothes. *Economic Journal* 99(supplement), 126–139.

Kirman, A. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6(2): 117–136.

Kiyotaki, N. 1988. Multiple expectational equilibria under monopolistic competition. *Quarterly Journal of Economics* 103: 695–713.

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

Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. New York: Oxford University Press.

Lindbeck, A., and D. Snower. 1987. Efficiency wages versus insiders and outsiders. *European Economic Review* 31: 407–416.

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

Lucas, R. 1978. Unemployment policy. *American Economic Review* 68: 353–357.

Lucas, R. 1987. *Models of business cycles*. Oxford: Basil Blackwell.

Lucas, R., and L. Rapping. 1969. Real wages, employment and inflation. *Journal of Political Economy* 77: 721–754.

Mankiw, N. 1985. Small menu cost and large business cycles: A macroeconomic model of monopoly. *Quarterly Journal of Economics* 100: 529–537.

Muth, J. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Nelson, A. 1984. Some issues surrounding the reduction from macroeconomics to microeconomics. *Philosophy of Science* 51: 573–594.

Pearce, D. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52: 1029–1050.

Phelps, E. (ed.). 1970. *Microeconomic foundations of unemployment and inflation theory*. London: Macmillan.

Roberts, J. 1987. An equilibrium model with involuntary unemployment at flexible competitive prices and wages. *American Economic Review* 57: 856–874.

Schelling, T. 1978. *Micromotives and macrobehavior*. New York: Norton.

Weintraub, E. 1979. *Microfoundations*. Cambridge: Cambridge University Press.

Yellen, J. 1984. Efficiency-wage models of unemployment. *American Economic Review* 74: 200–205.

Young, H. 1993. The evolution of conventions. *Econometrica* 61: 57–84.

Military Expenditure

R. P. Smith

In 1983, total world military expenditure was estimated to be $800 billion, about 30% higher in real terms than 1974 (SIPRI 1984). Although the bulk of the expenditure was by industrialized countries, the fastest growth was among the poor countries of the Third World. The UN study on the relationship between disarmament and development estimated that in the world as a whole in the late 1970s, military expenditure accounted for about 6% of world GNP and employed about 50 million people, including 25 million in the armed services, 5 million in the defence industries, and half a million scientists and engineers working on military R&D.

These numbers are very approximate, because of the deficiencies in the data on military expenditure. Many countries are secretive or misleading about their budgets. The treatment of items such as para-military forces and veterans pensions raises definitional problems. The line between civilian and military developments in space, nuclear and other industries, is ambiguous. Expenditures may not correspond to true economic costs, in particular where there is conscription. Comparisons over time raise problems about the choice of appropriate prices, while comparisons between countries raise exchange rate difficulties.

Economists have waxed and waned in their concern with military expenditure. The emphasis has varied depending on whether it is regarded as no different from other types of government expenditure or whether it is treated as having special features and a distinctive economic impact. However, in the process a large though disparate literature has grown up, linked to related work in international relations, peace studies and political science. Economists have also played an important, though controversial, role in the technical development of strategic doctrine, deterrence theory, nuclear targetting and other aspects of the uses to which military expenditure is put.

Because of its importance in the Government budget, the examination of military expenditure was an object of early attention by economists. The works of Adam Smith, Ricardo and Malthus contain relatively sophisticated discussions of issues involved in the provision and structure of military forces, their financing and their effects on aggregate demand. That constant in discussions of military topics, the cost escalation of weapons as result of new technologies, also makes an early appearance. Kennedy (1983, ch. 1) reviews the history of the analysis of military expenditure.

Historically, the concern of the British classical economists with military expenditure was understandable. Kohler (1980) estimates that during the century before 1815 the share of military expenditure in British GDP was over 10% during the War of the Austrian Succession, the Seven Years War, the US War of Independence and during the Napoleonic Wars, when at times it took over 20% of GDP and had a major negative impact on British industrial growth. From 1815 to 1914 military expenditure never accounted for more than 10% of British GDP. During the interwar period, economists, including Pigou, wrote widely on the issues of disarmament, but by 1939 attention had largely switched to organizing and paying for World War II, during which military expenditure took over half of British GNP. During the 1950s and 1960s, studies of military expenditure flourished, particularly in the US, along distinct macroeconomic and microeconomic strands.

The macroeconomic strand emphasized the domestic function of military expenditure in maintaining demand within capitalist societies prone to 'underconsumption' and stagnation. Baran and Sweezy (1966) is the most influential exposition of this interpretation, which seemed to be supported by the apparent efficacy of re-armament in curing the mass unemployment of the interwar period together with the postwar combination of historically high peacetime levels of military expenditure with low levels of unemployment. This argument is criticized in Smith (1977).

The emphasis on the economic utility of military expenditures prompted concern about whether disarmament was feasible and a range of empirical studies were carried out; Leontief

et al. (1965) for the US and Leontief and Duchin (1983) for the world are classics. The general conclusion of these and other studies has been that, contrary to the underconsumption argument, there seem to be no economic obstacles to disarmament. Nonetheless, the dispute as to whether military expenditure is an essential economic prop or a damaging burden persists. The economic effects of military expenditure in the US are reviewed by DeGrasse (1983) and in the UK by Chalmers (1985), both concluding that high military expenditure has played an important role in relative economic decline.

The microeconomic strand of research flowed from the wartime work on military planning done by economists using optimizing techniques. The techniques suggested for the efficient management of defence budgets, largely developed at the Rand Corporation and summarized in Hitch and McKean (1965), were implemented by McNamara's 'whiz-kids' in the US Department of Defense during the 1960s. The impact of this approach on Pentagon decision-making remains controversial, but the techniques introduced now play a major role in the theory of public policy appraisal. The three main questions posed within this framework all acquired pet names. Questions about 'how much is enough?' (the appropriate level), 'guns and butter' (the opportunity cost) and the 'bang for a buck' (cost-effective weapons procurement) still recur in the literature.

Central to any analysis of the level of military expenditure is some explanation of the forces which determine it. Different theories emphasize different factors: domestic militarism or international strategy, economic needs or political pressures, rational decisions or vested interests. The microeconomic strand, which still constitutes the dominant approach within defence economics, adopts, whether for explanatory or advisory roles, the perception of the state as a rational actor that balances the security benefits of the forces acquired against the opportunity costs of civilian expenditures foregone in order to determine 'how much is enough'. Against this, others emphasize the ability of the Military Industrial Complex to determine and shape defence decisions. Despite efforts, such as Rosen (1973), to test the competing theories there is little prospect of a resolution of these disputes.

There has been a wide range of empirical work on the economic impact of military expenditure, often measured in terms of the 'military burden' – the share of military expenditure in GDP. The only general conclusion seems to be that while the economic impacts are important, because of the size of these expenditures and their volatility, the precise effect is contingent on other factors. The effect will depend on the nature of the change in military expenditure, the economic circumstances prevailing, the social context of the programme, and the strategic environment of war or peace. Overall, there does not appear to be any systematic relation between military expenditure and unemployment, inflation, or the balance of payments. In each historical case the observed relation is the outcome of a number of effects operating on supply and demand in different ways.

One common, though not universal, result is that there is a trade-off between the shares of military expenditure and total investment in output. This relation seems to play a role in the negative association between the growth rate and the military burden across OECD countries. It is often claimed that there are technological 'spin-offs' from military R&D to civilian innovation, but the evidence for such positive effects on growth is scanty and contradictory. The technological impact is important, since in the mid 1980s defence absorbed over half the public Research and Development Budget in the US and UK, though a much smaller proportion in other western countries.

Within the Third World, rather different issues have arisen. Military forces appear to be needed to ensure autonomy from imperialist powers; to maintain external security in the face of deep-rooted regional antagonisms; to preserve internal unity against the divisive pressure of domestic conflicts; and to provide symbols of national status and prestige. It is also argued that the military is a technocratic modernizing force transforming economic, political and social relations within the

country. There is a considerable econometric literature on whether there is a positive or negative association between the military burden and the growth rate in Third World Countries, which is reviewed in Deger (1986).

Whatever the strength of these justifications for military expenditure, the consequence has been that tens of millions have died in the hundreds of wars in the Third World since 1945; military governments have become the norm; many poor countries have gone heavily into debt to purchase modern arms from the industrialized world; and some countries, such as Brazil, have begun to acquire substantial arms industries.

See Also

- ▶ Arms Races
- ▶ Defence Economics
- ▶ War Economy

Bibliography

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

Chalmers, M. 1985. *Paying for defence: Military spending and British decline*. London: Pluto.

Deger, S. 1986. *Military expenditure in third world countries*. London: Routledge & Kegan Paul.

DeGrasse, R.W. 1983. *Military expansion, economic decline: The impact of military spending on US economic performance*. Armonk: M.E. Sharpe.

Hitch, C.J., and R. McKean. 1965. *The economics of defense in the nuclear age*, 2nd ed. Cambridge, MA: Harvard University Press.

Kennedy, G. 1983. *Defence economics*. London: Duckworth.

Kohler, G. 1980. Determinants of the British defence burden. *Bulletin of Peace Proposals*.

Leontief, W., and F. Duchin. 1983. *Military spending*. Oxford: Oxford University Press.

Leontief, W., et al. 1965. The economic impact – Industrial and regional – Of an arms cut. *Review of Economics and Statistics* 47(3): 217–241.

Rosen, S. (ed.). 1973. *Testing the theory of the military industrial complex*. Lexington: Lexington Books.

SIPRI. 1984. *World armaments and disarmament: SIPRI yearbook 1984*. London: Taylor and Francis for the Stockholm International Peace Research Institute.

Smith, R.P. 1977. Military expenditure and capitalism. *Cambridge Journal of Economics* 1(1): 61–76.

Mill, James (1773–1836)

Donald Winch

Abstract

James Mill, Indian civil servant, Benthamite, and father and mentor of John Stuart Mill, introduced Jean-Baptiste Say's law of markets into British economic discourse. In addition to important works on the history of India, political and legal reform, and associationist psychology, he was the author of a textbook of Ricardian economics and played a major part in convincing Ricardo that he should write his *Principles of Political Economy* (1817). Through his son he was also responsible for giving prominence to proposals for taxing the 'unearned increment' in rental incomes that were influential in forming radical and socialist thinking in Britain.

Keywords

Associationist psychology; Bentham, J.; Betterment charge; Birth control; Corn Laws; Deductive method; Effective demand; Fabian economics; Labour theory of value; Land nationalization; Land tax; Law of rent; Macauley, T. B.; Malthus, T. R.; Methodology of economics; Mill, J.; Mill, J. S.; Philosophic radicalism; Productive and unproductive labour; Ricardo, D.; Say, J.-B.; Say's law; Stewart, D.; Underconsumptionism

JEL classifications

B31

Mill was born in a village near Montrose in Scotland, the son of a cobbler-cum-smallholder.With the support of a local laird, Sir John Stuart, he was able to attend Montrose Academy and then, in 1790, Edinburgh University, where his original goal was to become a minister in the Scottish Kirk. During the seven years he spent in

Edinburgh, he appears to have virtually become a member of the Stuart family, acting as tutor to the daughter of the house. Mill attended Dugald Stewart's lectures on moral philosophy and may have attended his class on political economy as well. Mill obtained his MA in 1794 and acquired a licence to preach in 1798. After an unsuccessful spell as an itinerant preacher and tutor, he moved to London in 1802, where he became part of an expatriate community of young Scots attempting to make their way in the world through journalism. In addition to various freelance jobs, Mill edited the *Literary Journal* from 1803 to 1806, writing most of the articles dealing with political and economic topics. This enabled him to marry in 1805 and begin a family of nine children that was to prove a strain on his finances and temperament. He also began work on what was to be an 11-year enterprise, the research for and writing of his *History of British India* (1817). In addition to hisincome from journalism, Mill obtained assistance from Jeremy Bentham, whose disciple and intermediary with the world of affairs he became after 1808. In this way, and especially through his articles for the Supplement to the 4th, 5th and 6th editions of the Encyclopaedia *Britannica* (1815–24), Mill became the leading light of the movement known as philosophic radicalism, an intellectual grouping dedicated to the reform of parliament and other legal and political institutions according to Benthamite criteria for 'good government'. In contradistinction to Bentham, however, Mill was a mature devotee of associationist psychology, as can be judged from his *Analysis of the Phenomena of the Human Mind* (1829). Mill also provided his eldest son, John Stuart, with an education which became part of the father's claim, both positive and negative, to have formed his son's mind and character. In 1819, partly as a result of the reception given to his *History*, Mill was appointed to the post of Assistant Examiner with the East India Company, rising to the post of Chief Examiner in 1830, a position he held until his death in 1836.

Mill's early economic writings consist of a large body of articles and two pamphlets, the first of which was *An Essay on the Impolicy of a Bounty on the Exportation of Grain* (1804), constructed along Smithian lines, the second entitled *Commerce Defended*: *An Argument by which Mr. Spence, Mr. Cobbett, and others have attempted to prove that Commerce is not a source of* National *Wealth* (1808). The latter is of interest to the history of economics, for two main reasons. The work contains the first enunciation in English of what was originally known as the Say–Mill law of markets; and it was through this work that Mill made the acquaintance of David Ricardo. The pamphlet was an attack on the views of those neo-physiocratic authors who argued during the period of Napoleon's economic blockade that agriculture rather than manufacturing and commerce was the true source of Britain's wealth. Mill agreed that claims on behalf of commerce had frequently been pitched too high, but he defended the Smithian view that manufacturing and other profits were a legitimate form of net surplus. He also upheld a pre-comparative cost interpretation of the gains from trade judged by the difference between the real costs incurred in producing goods for export and the putative domestic cost of producing imported goods. In countering Spence's underconsumptionist arguments on the relationship between capital accumulation, consumption and public expenditure, Mill defended Smith's distinction between productive and unproductive labour, translating it into the goods consumed by each category in order to show the importance of accumulation and productive consumption to economic growth. In refuting the idea of excessive accumulation, or general overproduction, Mill invoked Say's principle: 'The production of commodities creates, and is the one and universal cause which creates a market for the commodities produced' (1808, p. 135). Since the argument was conducted in barter terms, however, it amounts to little more than a statement of Say's identity, though the implication was that the conclusions applied equally to a money economy. Hence Mill's conclusion: the claims of commerce could be exaggerated whenever it was suggested that the extension of foreign markets was necessary to guarantee full employment. Here then was the English origin of the idea, expressed in characteristically unqualified terms, that was to lie at the

heart of the controversy between Ricardo and Malthus over general gluts, and was later to be taken up by Keynes as the distinguishing mark of orthodox classical (and neoclassical) macroeconomics – an intellectual obstacle that had to be removed by a new theory of effective demand in order to open the way for an explanation of involuntary unemployment in the *General Theory.*

It was largely as a result of Mill's encouragement that Ricardo overcame his doubts as to his capacity to move from being an economic pamphleteer to writing his *Principles of Political Economy*, which embodied those new doctrines that were necessary in order to replace those of Smith and other predecessors. Mill became Ricardo's impresario, coach and disciple; he was responsible for completing Ricardo's education and inducing him to enter parliament as spokesman for the 'true' principles of political economy and the reform programme of philosophic radicalism. Mill wrote one of the first 'schoolbook' accounts of Ricardo's doctrines in his *Elements of Political Economy* (1821), a record of what his son was taught at the tender age of 13. Ricardian doctrines appear in their most simplified and abstract form, but arranged according to the model provided by Jean Baptiste Say's *Traitéd'économiepolitique* (1814), and with some embellishments that were not always acceptable to Ricardo himself. Thus in attempting to defend Ricardo's labour theory of value from attack by Robert Torrens, Mill bowdlerized the theory.

On policy matters, however, Mill struck out more boldly than Ricardo on two main issues: the advocacy of birth control as a solution to the problem of low wages, and a proposal for taxing the increment in rents accruing to landowners as a result of any legislative action which increased the demand for land. (Mill chiefly had the Corn Laws in mind.) He was sympathetic to land nationalization (if only as a way of frightening the landed aristocracy) and to the view that taxes on rent were one of the best means of raising government revenue; but he recognized that such proposals could not be introduced into a country where property had already exchanged hands at prices reflecting rental expectations. Nevertheless, since this only gave a legitimate expectation to present rents, plus an allowance for improvements undertaken by the landowner, Mill was in favour of levying what would later be called a 'betterment' charge on increments in rent beyond this. In adopting this position he believed he was merely carrying Ricardo's conclusions on the special nature of rent, as compared with wages and profits, to their logical policy conclusion.

Mill then, rather than Ricardo, is the source of that strand of radical thinking on the 'law of rent' that was to be passed on via his son to the Fabians. More significantly, when judged by results, the official positions occupied by Mill and his son in India House ensured that their views on taxation and land revenue were influential in practice. It was primarily through his efforts that a determined attempt was made in the Bengal provinces to replace a landowner-based *(zemindari*) system of land tenure with one based on the view that the government should retain the ultimate rights in land and deal directly with the peasant cultivator or ryot, basing the tax assessment on Ricardian or pure rent.

Mill is also of some importance for his views on the methodology of political economy and other moral sciences, as can be best illustrated – negatively at least – by the attack mounted by Macaulay on Mill's essay on 'Government' for the *Encyclopaedia Britannica*. Mill was an extreme upholder of the virtues of the deductive method, and a critic of practical men who professed to be 'all for fact and nothing for theory'. In this respect Mill is sometimes credited with an influence on, certainly as encouraging, Ricardo's adoption of the a priori method of working from unqualified assumptions to 'strong cases', and from there to policy conclusions. Since there is little firm evidence to establish this proposition, those who are either critical or defensive of the Ricardian method should probably dispense with Mill rather than attempt to draw attention to similarities or differences between the practice of both men. We do know, however, that Mill produced a son who believed that his education had peculiarly fitted him to engage in 'the science of science itself, the science of investigation – of method'. We also know that in the aftermath of the

Macaulay attack the son wrote an essay 'On the Definition of Political Economy; and on the Method of Investigation Proper to it', later to be the basis for Book VI of his *System of Logic* (1843), in which he criticized both his father and Macaulay in the course of expounding an interpretation of the role of deductive methods in political economy which remained canonical for much of the 19th century.

See Also

- ▶ Classical Distribution Theories
- ▶ Enlightenment, Scottish
- ▶ Land Tax
- ▶ Mill, John Stuart (1806–1873)
- ▶ Ricardo, David (1772–1823)
- ▶ Say, Jean-Baptiste (1767–1832)

Selected Works

1804. *An essay on the impolicy of a bounty on the exportation of grain*. London. Repr. in Winch (1966).
1808. *Commerce defended: An argument by which Mr. Spence, Mr. Cobbett, and others have attempted to prove that commerce is not a source of national wealth*. Repr. in Winch (1966).
1817. *History of British India*. London: Baldwin & Cradock.
1821. *Elements of political economy*. London: Baldwin Cradock & Joy; 2nd ed, 1824; 3rd ed, 1826. Repr. in Winch (1966).
1829. *Analysis of the phenomena of the human mind*. London: Baldwin & Cradock.

Bibliography

Bain, A. 1882. *James Mill: A biography*. London: Longmans.
Ball, T. 1992. *James Mill: Political writings*. Cambridge: Cambridge University Press.
De Marchi, N.B. 1983. The case for James Mill. In *Methodological controversy in economics: Historical essays in honor of T.W. Hutchison*, ed. A.W. Coats. Greenwich: JAI Press.
Keynes, J.M. 1936. *General theory of employment interest and money*. London: Macmillan.
Macaulay, T.B. 1829. Mill's essay on government: Utilitarian logic and politics. *Edinburgh Review* 97.
Mill, J.S. 1843. *A system of logic*. London: Parker.
Mill, J.S. 1844. On the definition of political economy; and on the method of investigation proper to it. In *Essays on some unsettled questions in political economy*. London: Parker. [Written 1829; first published in the London and Westminster Review, 1836].
Ricardo, D. 1817. *Principles of political economy and taxation*. London: John Murray.
Say, J.-B. 1814. *Traitéd'économiepolitique*, 2nd ed. Paris: A.-A. Renouard.
Winch, D. 1966. *James Mill: Selected economic writings*. Edinburgh: Oliver & Boyd. Reissue, New Brunswick: Transaction, 2006.

Mill, John Stuart (1806–1873)

N. De Marchi

Abstract

Mill approached economic theory using conceptual and verbal analysis. This worked well for settled truths applied to circumscribed situations, such as a rise in the ratio of food prices to manufactured goods prices under growth subject to decreasing returns. He needed, but did not develop, a different method for multicausal problems. Mill insisted that value and production were settled areas of political economy but was open to societal reforms that would result in altered shares of income and wealth. This distracted from the coherence of his *Principles of Political Economy* and from his reputation as a theorist, while ensuring that he will be remembered for challenging readers to entertain breathtaking prospects for human improvement.

Keywords

Absolute and exchangeable value; Bentham, J.; Birth control; Communal ownership; Corn Laws; Education; Falling rate of profit; Happiness; Inheritance and bequests; Labour as a measure of value; Laissez-faire; Mill, J.; Mill,

J.S.; Rent; Ricardo, D.; Role of government; Smith, A.; Social preferences; Speculation; Stationary state; Taylor, H.; Value

JEL Classifications
B31

John Stuart Mill was the pre-eminent British economist of the mid-19th century. But he was much more besides, commanding a hearing in public debates on subjects from logic to liberty, the position of women to the problem of Ireland. Yet, though his *Principles of Economics, with Some of Their Applications to Social Philosophy* (1848) dominated economic discourse for 40 years, there is little in it of technical, or even conceptual, advance that would justify placing him in a pantheon of great economists, if one judges by what is understood as economics today. Mill should be known and honoured more for his vision of an improved condition for humankind and for the novel economic views that formed part of that vision than for his economic analysis as such.

Approaching Mill's economic ideas from this perspective necessitates attending to his passage from early Benthamite propagandist and defender of Ricardian doctrine to more pragmatic reform strategist, with a greatly expanded notion of happiness. The transformation was traumatic in that it involved a lapse, and relapses, into depression, and it meant modifying some old convictions. Positively, however, Mill also discovered the possibility of cultivating feelings – 'of inward joy, of sympathetic and imaginative pleasure' – and began to see for others the prospect of 'perennial sources of happiness, when all the greater evils of life shall have been removed' (1873, pp. 147, 151). Certain elements in this prospect seemed to him to require, ultimately, the replacement of competition with cooperation, and there were various other novel economic aspects to this notion. But the inspiration was quite different from the motivations reflected in the economics Mill had learned from James Mill (his father) and Ricardo. This makes it hard to find the strong logical link between economic doctrine and social philosophy implied by the word 'Applications' in the subtitle of his *Principles*. In fact there is a switch in mode between the doctrines and the applications, from the demonstrative to the conditional – from result to possibility – which seriously weakens that link. There is also a difference of tone: Mill wrote with great immediacy and verve about possibilities for the improvement of humankind, but defensively on the economic doctrines he had inherited. He embraced new social thinkers, borrowing freely from them even if, as he often allowed, their views were incomplete, not always coherent and even downright misleading in certain respects. But he chose not to keep up with new developments in economics, more especially those that employed mathematics. Instead he stuck to the method that was his forte, clearing up terminological and logical confusion and thus 'perplexities'.

The Constraints of a Benthamite Education

Mill was the eldest of eight children born to James Mill and Harriet (née Barrow). He was homeschooled by his demanding father, whom he eventually succeeded as Examiner in the East India Company. The elder Mill was a Scots literary émigré in London, disciple of Bentham, and a leading protagonist of utilitarian reform. His writing took precedence, and, besides giving John basic instruction, he largely turned over to him the education of the younger children. John's mother, worn down, developed no intellectual interests, and became for him a model of what women should not be, in sharp contrast to Harriet Taylor, with whom he fell in love in the 1830s and married in 1851. Harriet Taylor shared Mill's reformist ideas and emboldened him in expressing his notions concerning autonomy, not least for women.

John's spectacular childhood achievements are well known: beginning to learn Greek words at the age of three, and starting Latin at eight, acquiring the language by dint of having to instruct his siblings. Studies in Logic began at 12, and Political Economy at 13. Instruction in the latter took

the form of lectures from his father, which he was to summarize and repeat the following day, on their daily walk. James Mill's *Elements of Political Economy* (1821), which the daily lectures became, was essentially a set of logical propositions. John always regarded logical analysis as the most valuable of all mental trainings.

At the age of 20, however, Mill discovered that something was lacking. In describing what he would later call a crisis in his mental history, he recalls imagining the accomplishment of all the Benthamite reforms for which he was agitating, but finding himself without satisfaction at the prospect. Recovery was effected, he tells us, through reading new authors and modifying his circle of friends; central to the process, however, was the realization that the Benthamite views he had imbibed were entirely too narrow.

During Mill's childhood the family spent their summers close to Bentham, and in 1821 he began reading him – in fact Dumont's edited version of notes, published as the *Traité de Législation*. This work gave him 'a vista of improvement' for human life based on coherent laws and opinions founded on the principle of utility (1873, pp. 69, 71). Somewhat later Mill was given the task of editing Bentham's manuscript of *The Rationale of Judicial Evidence* (1827); so, by the time of his depression, Mill must have been as well equipped as anyone in Britain to convey accurately Bentham's thinking on government and to comment on matters of English law, which he did, frequently, in the daily and periodical press.

Mill's post-depression reappraisals were cautious, even after Bentham and his own father had died. In 1838, however, he was able to present a lengthy and balanced account. He praised Bentham for having accomplished the first scientific investigation of the large and messy body of precepts that comprised English law, using as his tool what Mill called 'the method of detail'– separating wholes into their parts, resolving abstractions into concrete things – in short, 'breaking every question into pieces before attempting to solve it' (1838, pp. 83, 100). The method was Baconian; Bentham's originality lay not in having invented it, but in having applied it to the law. He had not yet had the impact that he deserved, partly because of his obsessive verbal partitioning of every topic. This resulted in tedious intricacies for which few readers cared. But the method had the great merit of bringing into question even commonly accepted truths and constituted a tool for identifying the rationale, or lack of it, in every existing or proposed law.

Take murder, for example. According to common sense and religion it is a crime. But why? A rational examination would ask whether the benefits to the perpetrator were outweighed by costs, in terms of the suffering inflicted on the victim; the feelings of insecurity aroused in others; and the discouragement to certain sorts of industry and useful pursuits through fear, as well as any diversion of resources to warding off the perceived danger. If the costs dominate, then murder must count as a crime and the infliction of punishment is warranted (1838, p. 83). Mill judged it useful to challenge even basic truths in this way, both because they support many subsidiary truths, and for the mental discipline involved, which we need in order to guard us against too readily following moralists who invoke, unexamined, phrases such as 'law of nature' or 'right reason', and politicians who call for 'liberty' and 'social order' (1838, p. 84; 1873, p. 67).

On the negative side, Mill found Bentham's approach cripplingly narrow. By focusing on pain and pleasure exclusively Bentham implied that human beings are governed solely by their own immediate interest and their sympathy or antipathy towards others.

Among things ignored are a feeling of moral approbation or disapprobation (conscience); standards of excellence or the desire of perfection as an end in itself; a sense of honour or personal dignity; a love of beauty; the passion of the artist; love of the congruency or consistency of things, or of their conforming to their intended ends (1838, pp. 95–6). This philosophy, devoid of morality and spiritual interests, did not sit well with the new Mill, who had now 'learnt by experience that the passive susceptibilities needed to be cultivated as well as the active capacities' (1873, p. 147). But neither was Bentham's philosophy able to cope well with even the purely business aspects of life,

since in practice every action influences our own and others' affections and desires (1838, p. 98).

Mill's own revised aspiration was to give due place to the *moral, aesthetic*, and *sympathetic* aspect of every human action. We must ask of each, is it right or wrong? Is it beautiful – inspiring, estimable? And is it 'loveable?' (1838, p. 112). These additions could have come from Smith's *Theory of Moral Sentiments* (1759), though it is not clear that Mill knew that work.

Returning to Bentham's philosophy, Mill concluded that, since it ignored feelings and the moral, the inspiring and the lovable, it simply offered a crude guide to desirable outward circumstances and regulations to effect them. But circumstances and punishments cannot instil the sympathy that binds us. Mill drew from the aftermath of the French Revolution the lesson that social feelings are only shallow-rooted in human nature, that, once a society has torn down old institutions that have grown corrupt, conflicting interests are likely to produce anarchy (1838, p. 99). For sympathy to prevail, then, there must be education directed towards making it second nature to care for others as we care for ourselves.

On the political level, Bentham, it was true, had urged that government be delegated to those whose interests are identical with the interests of the population at large. But Mill feared giving even such a group control over the whole; without a serious opposition its members are apt to become tyrannical (1838, pp. 106–8). Mill's fears in this regard presage those expressed by Hayek in his *Road to Serfdom* (1944).

What did it mean to be a Benthamite propagandist, as Mill was before 1826? Two examples will illustrate. He distributed pamphlets on methods of birth control, convinced that the average condition of the working classes could be permanently improved only by voluntary reduction in their numbers relative to the means available for their support. And he opposed the Corn Laws because, in restricting imports, they kept the price of grain higher than it need be, making the most basic means of sustenance less accessible, which was a clear net loss of aggregate happiness. By contrast with these cut-and-dried policy choices, Mill's views in the decade or so after 1826 were largely an outworking of the enlarged basis for personal and social happiness that he had adopted.

He also became more practical; he saw that radicals must co-opt conservatives to command a parliamentary majority. The new Mill judged that there is no simple and direct connection between first principles, such as the principle of utility, and actions that will increase happiness. For individuals differ in their primary beliefs, making happiness 'too complex and indefinite an end' to pursue in the Benthamite manner (1838, p. 110). Fortunately, division on ultimate standards does not preclude agreement on intermediate ends. During the 1830s, therefore, Mill strove to engage erstwhile opponents on such intermediate goals, arguing, for example, that the landed interest's support for the Corn Laws would be weakened if it could be shown that those laws actually increased wage costs, harming landowners both as employers and consumers.

Early Political Economy

Philosophically and in terms of political practice Mill's new views had far-reaching implications for his life and writing. Much of his political economy, however, underwent relatively little change. Not only prior to 1826, but even in his *Principles* he retained and defended the core doctrines of his father's *Elements* and Ricardo's *Principles of Political Economy and Taxation* (1817). At times, especially early on, his defence was conducted with a fierceness that blinded him to any merit in alternative views. In the 1840s, by which time he wanted to make place, alongside the old core doctrines, for many additional topics in economic analysis as well as for his favourite ideas for improving society, the combination lent to his *Principles* the appearance of a patchwork. An illustration of his early treatment of critics can be given here; the patchwork aspect of the *Principles* will be touched upon in section "Mill's Mature Political Economy".

The illustration concerns the central subject of value. Ricardo and James Mill chose to discuss value strictly in terms of exchange value, for both

wanted to show that, as population grows, the value of food *relative* to manufactures rises because land of lesser productivity has to be brought under cultivation. On the assumption that returns in manufacturing are constant but unit labour costs in agriculture are rising, it is obvious what causes an observed rising trend in the relative price of food.

Smith, however, in addition to allowing value to be relative, stressed the pain cost of labour and insisted that the true cost of goods is how much labour they command. Behind this emphasis was a concern that the sacrifice of ease involves a loss of happiness, since ease for Smith was linked with tranquillity of mind, and the latter with happiness. Mill understood this (see 1848, pp. 580–1). Nevertheless, as a young defender of his father and Ricardo, he dismissed Smith's alternative measure of value lest readers be deflected from focusing on relative labour input, so central to the case against the Corn Laws. Hence, when Malthus, in his *Measure of Value* (1823), opted for Smith's sacrifice measure, Mill, aged 17, portrayed him as logically incompetent: to make labour command a measure of value, Malthus had in fact to assume what he needed to show, that the value of wages is always the same (1823, p. 57). Mill was correct but also quite one-eyed. Malthus showed in his *Definitions in Political Economy* (1827) that he too grasped the difference between an invariable measure of value and exchange value, yet preferred to measure even exchange value by how much labour commodities can command because that is appropriate if one's purpose is to ascertain 'the sacrifice which people are willing to make in order to obtain [commodities]' (1827, p. 211).

Against this crabbed performance, it is refreshing to find Mill, a very few years later, writing comparatively wide-ranging and subtle analyses of current events. The best of these was an essay, 'Paper Currency and Commercial Distress', in the short-lived radical *Parliamentary Review* of 1826, on the recent 'commercial revulsion'.

Mill insisted that the proximate cause of recession in this case was a prior speculation, not in new ventures, but in existing activities. The dominant group of parliamentarians instead blamed an over-issue of small notes by country bankers – an attribution of causation, Mill suggested, that betrayed a deeper scepticism about paper currency. Drawing on Tooke's recently revised *Considerations on the State of the Currency* (1826), Mill showed that the speculation had begun after trade papers reported below normal stocks in a few key goods, including grains. In the usual way this had induced dealers to increase their purchases, causing an immediate price increase, a pattern that then extended itself, though for purely speculative reasons, to a wider range of goods. Mill agreed that there had been an increase in credit associated with speculative buying, but observed that this did not require small notes: trade credit and bills of exchange would have sufficed. He also showed that the observed movements in the currency were not what one would expect from an expansion of the circulation. What had happened was merely a redistribution of, rather than an overall expansion in, the circulation. When grain prices first began to rise, means of circulation shifted from London to the country, sustaining the rise in agricultural prices but lowering the prices of manufactures in the city. Manufactured exports therefore rose, and, because grain imports were restricted under the Corn Laws, the exports occasioned an influx of gold. This would have happened whether the medium of circulation was metallic or paper, and no net expansion of their notes by country bankers need have been involved. It followed that the ultimate culprit was the Corn Laws, which prevented imports from offsetting the speculative purchases occasioned by the initial shortfalls in stocks of grain.

This was a tour de force in applied economic analysis. Mill contrasted his analysis with an account of why the parliamentarians had got it wrong. At root they lacked general principles. Inevitably, then, the views of 'practical men' – men who observed a few facts near at hand and generalized on that inadequate basis – prevailed with them.

Practical men as nemesis was a theme in Mill's famous essay 'On the Definition of Political Economy; and on the Method of Investigation Proper to It', which was published in 1836 and again, along with other youthful exercises in clarification

of the principles of the new political economy, in *Essays on Some Unsettled Questions of Political Economy* (1844). From an economist's point of view perhaps the most useful portion of Mill's *System of Logic, Ratiocinative and Inductive* (1843) was his extended analysis of the social scientist's equivalent of experimentation: the various methods of ascertaining causes (Book III). The earlier methodological essay started him down that road.

The methodological essay is by far the most sophisticated of the set published in 1844, the others, with two other exceptions, suffering from being of the crabbed, defensive sort. In an essay on 'The Influence of Consumption on Production' Mill allowed that a general glut could occur, temporarily, if there were a sudden general preference for liquidity. The other exception was an essay on 'The Laws of Interchange between Nations', in which Mill elaborated on his father's suggestion that the division of the gains from trade would depend on the relative strengths of demand of the participating countries. This was one of Mill's few lasting contributions to economic analysis. Marshall utilized it in his essay *The Pure Theory of Foreign Trade* (1879), and his demonstration, in the context of the 1903 tariff debate that whether the foreigner bears the cost of a tariff will depend on the shape of his offer curve.

Espousing Selective Conservatism, and Incorporating Social Evolution

By the late 1830s, as we have seen, Mill had begun to make explicit what was required to make good on Bentham's omissions. But how exactly were these to be supplied? Here Mill had recourse to German views, conveyed in language more palatable to English minds by the poet and essayist Samuel Taylor Coleridge. He also drew on the writings of the Saint Simonians, particularly the early work of Auguste Comte.

Mill took from Coleridge the idea that education should assist in forming national character. The young need to be imbued with an 'active principle of cohesion', of sympathy, not hostility; union, not separation. This might require heroes, or at least common beliefs; either way the goal must be to make caring for others second nature. By implication, there was a very active role here for government, a role more positive than either the pre-revolutionary French philosophers had allowed, or than their English counterparts had felt to be necessary. The French had wanted to tear down corrupt and spent institutions, after which government should basically leave people be (laissez-faire). On the English side, the national discomfort with conflict and a preference for compromise had asserted itself in the 18th century; after the strife of the 17th century the English had settled for living with whatever institutions there were, provided they were reduced to practical nullities (1840, pp. 142–4, 146). There was no sense in England that education should be reformed to build national character and supply an active force for social cohesion.

Mill picked up on three intriguing ways in which government might contribute to or reflect cohesion; and each had an economic aspect. First, the state 'ought to be considered as a great benefit society, or mutual insurance company, for helping (under the necessary regulations for preventing abuse) that large proportion of its members who cannot help themselves' (1840, p. 156). The details of this idea were not filled in, and it does not reappear in Mill's later work, but it sounds not unlike the Social Security system of the United States or the mandatory contributions towards retirement now applied in many countries.

Second, the land must be considered a trust. Mill distinguished this notion from calls for the state to reclaim private property, though he noted that the law of real property originally applied only to movables. It was his view that, if an owner possesses more land than is necessary for him to sustain himself and his family by his own labour, the excess confers on him power over others and the state may require that this power not be abused. This meant that even the system of cultivation is a proper concern of society (1840, pp. 157–8). The notion reappears in the Principles, though as one among several possibilities for limiting bequest and tenure (1848, p. 227).

Third, Mill insisted that education, being of almost boundless power, should be used by the

state to foster public opinion in favour of the attracting forces within society. These forces derive from our love of praise, favour, admiration and respect, and our dread of shame and ill repute – again, ideas central to Smith's *Moral Sentiments*, though Smith was not acknowledged by Mill. Mill held that, once the basic means of living has been obtained, almost the whole of our remaining effort is directed to acquiring the favourable regard of others. In fact this is the driving force behind the industrial and commercial activity that advances civilization. Love of praise, however, is also the source of the selfish thirst for aggrandizement; hence the state must tip the balance in favour of social sympathy (1840, pp. 410–1).

A possible explanation of Mill's slighting of Smith is available in this instance. Mill might easily have seen Smith as insufficiently positive about the role of government. Smith, for example, advocated basic education for the poor in the hope that, for those condemned by excessive specialization to repetitive, trivial tasks, it might mitigate the risk of moral deformity (1776, p. 788). But for Mill that was too feeble a response, too restrained an expectation. For him education was the key to all future social and personal improvement.

The expectation of improvement also impelled Mill farther in a related direction. There is an implication for distribution in the notion that mutuality of interests makes it easier to cultivate and fix social feelings (1861, p. 231). Mill took from Comte the conviction that there had been considerable social progress towards cooperation, a trend likely to continue. The cooperative spirit, in turn, ought to make it possible for individuals to regard working for the benefit of others as a good it itself, requiring no compensation. Ideally, what we get for ourselves should not be viewed as a quid pro quo for our cooperation but in terms of 'how much the circumstances of society permit to be assigned' to us, 'consistent with the just claims of others'. The market method of settling a worker's share of the produce may be a temporary practical necessity, but morally is not ideal. Society, Mill understood, was not yet ready to relinquish the market, so he judged it better to let competition decide rather than to impose any artificial mode of distribution as yet untried – save in the army, where it was the de facto norm (1865, pp. 340–1). The idea reappears in the discussion, in the *Principles,* of cooperative arrangements in industry, though Mill's emphasis there was strongly on shared ownership for harnessing 'productive energies', and cooperation as still for the future, while competition is not only dominant but also has its positive side emphasized (1848, pp. 216, 337, 356).

Happiness: An Enlarged View

Feelings aside, morals, aesthetics and sympathy – the other three missing elements in Bentham's philosophy – put happiness firmly in the social sphere. Mill continued to hold, with Bentham, that the general end should be the multiplication of happiness, but increasingly happiness had to involve the desire to care for others. Even if by nature we have only a small germ of this feeling it is one which can and should be 'laid hold of and nourished by the contagion of sympathy and the influences of education' and supported by external sanctions (1861, p. 233). Social ends would thus be rendered part of our inmost motivation.

On the one hand, then, Mill naturally came to think of happiness as linked to the growth of the cooperative spirit. On the other, he also saw it, crucially, as involving the development of the inward man, which is where the three added dimensions really have their purchase on our emotions and motives. He would eventually redefine individual happiness as a satisfied life, one with a balance between tranquillity and excitement. A person who finds this balance can be content with little pleasure, and can even be reconciled to much pain (1861, p. 215). Mill saw no reason why the mass of humankind could not unite tranquillity with excitement, since, even without great improvement in outward circumstances, the inward balance could be struck.

Notice, however, that inward happiness, since it does not depend essentially on a person's material resources or situation, removes the end – happiness – from the status of positional good. It may be that this realization predisposed Mill to

accept Comte's ideas on cooperation – that cooperation itself is made easier if the overall end in view does not involve rivalry – though there is no collaborating evidence for this.

Mill's Mature Political Economy

Mill's *Principles* was an uneasy amalgam of Smith, Ricardo, Mill's own refined insights on various discrete topics, and new social ideas.

The treatise can be dissected for its insights on a wide range of topics, as Hollander has done in his full-length study of Mill (1985). Hollander shows Mill to have had an unusually clear grasp of mechanisms: the determination of (long-run or cost) prices by variation of supply; of the rate of return by the proportion of the work day required to produce wage goods; of the alternative to wage reduction that exists in population control, as a way of equating the growth rates of population and capital accumulation; of the feedback between speculation engendered by a declining rate of profit and the loss of capital due to business failures, which in itself will raise the rate of profit; of a general desire to hold money as a cause of depressions; and so on.

These mechanisms summarize clearly and appropriately Mill's analytical contribution, which he even recorded on occasion as a list of propositions established (for example 1848, pp. 497–9). Since, however, there are various possibilities implicit in the application of such propositions, circumstances matter, as Mill himself stressed in his essay on method. This distinction between demonstrated truth and institutional possibilities inevitably loosened the logical connection between Mill's economic analysis and his social views. Thus he could analyse in a Malthusian-cumRicardian way the growth tendencies that issue in stationariness: given diminishing returns in agriculture, constant population growth and a fixed state of the productive arts, growth will eventually cease. Yet he could also freely explore possibilities for human nature, society and the 'Art of Living' in the stationary state, unconstrained by those economic tendency laws.

This is partly responsible for the patchwork appearance of the *Principles* noted earlier; yet it probably owes as much or more to Mill's having retained key doctrines from Ricardo and his father while accepting that they had not always elaborated the general case. In acknowledging this, on rent for example, Mill ended up incorporating qualifications that the reader must locate here and there – in the case of rent, in four separate chapters, spread across three books.

Marshall chose an alternative way of addressing rent. He noted that rent of land is 'no isolated economic doctrine' but 'simply the chief species of a large genus of economic phenomena' (1890, I, p. 629). Mill, sticking to the Ricardian view that rent of land is 'differential and peculiar' (1848, p. 495), concluded that rent only enters into the cost of production if there is a scarcity element involved – cases 'rather conceivable than actually existing' (1848, p. 498). Marshall, however, constructed a continuum of cases in which, at one extreme, a productive resource is in strictly fixed supply and its return therefore a surplus or 'rent in the strictest sense of the term', while at the other end the resource is quickly reproducible and its return no more than the interest on the money cost of obtaining more of it. There are multiple combinations in between where revenue might temporarily diverge from interest, for reasons originating either on the supply or the demand side. Specifying the exact circumstances may have consequences, as when a choice must be made whether to impose a tax on producers rather than consumers. Marshall's point was that interest and quasi-rent 'shade into one another gradually', making such choices very difficult (1890, I, pp. 412–21). Hence, as to 'rent not entering into cost', he concluded that the phrase cannot be rescued by verbal analysis but 'only by experience'. At the same time, it is a 'denial of subtle truths' to generalize either in the direction chosen by Mill or its opposite (1890, II, p. 439). Mill's fierce defence of Ricardian doctrine in this instance, as in some others, did not advance the cause of clarity nor did it allow experience the crucial role his own method suggested it should have.

As noted, Mill incorporated analytical developments in economics selectively; he left aside

those that involved mathematics – not the strongest component in his early education (1873, p. 15, though see also p. 59) and a mode of reasoning he later came to suspect of strengthening the false claim that moral, political and 'supersensual' truth may be had without self-observation and common experience (1832, p. 331; 1873, p. 233). Not to speak of French works, he failed to mention even contemporary English analyses of profit-maximizing equilibrium, and the gains and losses from the supposition of various changes (in technique – hence machinery – or in taxes), such as those due to Tozer (1838) and Lardner (1850). Much later he responded to Jevons, though probably not from having studied the *Theory of Political Economy* (1871) at first hand. And from reviews of the *Theory* Mill *mis*judged that Jevons just offered 'a notation implying the existence of greater precision in the data than the questions admit of' (Mill to Cairnes, 5 December 1871, in Mill, 1963–91, vol. 17, p. 1862).

There remain, as the freshest contributions of the *Principles*, those of Mill's notions on future social possibilities that have some economic content.

1. In the context of reflecting on possible distributions of property (Book II, Ch. 1), Mill posited that a society might be in the position of having to choose between communal ownership and private. He supplied arguments why the communal arrangement ought not to be rejected out of hand. Shirking, he allowed, would be a serious problem; moreover, the experiment should not be tried without universal education first being implemented and numbers (population) controlled, so that none would lack for subsistence. Under such circumstances one might assume more public spirit than we are used to seeing. Nevertheless, and difficulties with the alternative notwithstanding, he thought existing production arrangements far from ideal: in nine-tenths of cases there are principal–agent problems (not his terminology). All said, he suggested, the choice should turn on the most important issue of all: which system 'is consistent with the greatest amount of human liberty and spontaneity'? (1848, Book II, Ch. 1, p. 208).
2. In the very next chapter Mill argued for a distinction between the right of private ownership and the right to bequeath and inherit. On the one hand, the power to bequeath might be inconsistent with the permanent interest of the race; on the other, the essential principle of property – to assure to all what they themselves have produced – cannot apply to the raw materials of the earth. After universally agreed exceptions, Mill observed, where doubt is present the presumption should be against the owner (1848, Book II, Ch. 2).
3. In Book IV, Chapter 4, Mill adumbrated his own non-Smithian tendency for the rate of profits to fall. He accepted the tendency, but argued that it reflects not only the natural (Ricardian) consequences of the extension of cultivation, but also the progress of civilization. As people become more rational they also become more self-controlled, and find lower rates of interest and profits acceptable. Not only are rational people less apt to discount the future; they also save against contingencies even in the absence of any immediate need. In a more civilized world, moreover, risks are lower because the strong social spirit renders capital and wealth generally more secure.
4. Building on the arguments just listed, Mill was also able to contemplate a future with zero growth (the stationary state: Book IV, Ch. 6). Here he reiterated the theme that 'a population might be too crowded' for that solitariness and tranquillity so essential to depth of character. Quite apart from that, zero growth of course does not preclude 'improving the Art of Living'. And in any case, the social ideal cannot be the elbowing, crushing competition all around us. We should be able to get beyond the struggle for (relative) riches, so as to realize a state in which 'while no one is poor, no one desires to be richer, nor has any reason to fear being thrust back' (1848, p. 754).
5. A third chapter in Book IV, 'On the Probably Futurity of the Labouring Classes', expands on all this, but stresses the importance of making people more 'rational' by increasing

their independence, this by reversing the hiring– service relationship and replacing it with employer–employee associations (1848, p. 763). As so often, Mill qualified this sweepingly optimistic view with a pragmatic caution: competition need not be dispensed with; after all, cheaper goods come of it and labour must therefore benefit (1848, p. 794).

6. Finally, in Book V, especially Chapter 11, there is an exploration of laissez-faire, the general rule, and the 'large exceptions' to it that Mill also deemed necessary. The positive role of government should extend to education; the care of minors (from which category he was careful to exclude women); and a long list of cases where private initiative would be preferable if only it were not generally lacking for one reason or another. The list reads quite like the one Smith provided, of desirable projects for which no individual or small group can find the necessary financing; only Mill extended it beyond roads, harbours, canals, and so on, to hospitals, schools, colleges and printing presses (1848, pp. 944, 947, 950, 970).

See Also

- ▶ Competition
- ▶ Cooperation
- ▶ Property Rights

Selected Works

1823. Malthus's measure of value. In *Collected works,* vol. 23.
1826. Paper currency and commercial distress. In *Collected works*, vol. 4.
1836. On the definition of political economy; and on the method of investigation proper to it. In *Collected works*, vol. 4.
1838. Bentham. In *Collected works*, vol. 10.
1840. Coleridge. In *Collected works*, vol. 10.
1843. A system of logic, ratiocinative and inductive. In *Collected works,* vols. 7 and 8.
1844a. Of the laws of interchange between nations. In *Essays on some unsettled questions of political economy. In Collected works,* vol. 4.
1844b. On the influence of consumption on production. In *Essays on some unsettled questions of political economy. In Collected works,* vol. 4.
1848. Principles of political economy, with some of their applications to social philosophy. In *Collected works,* vols 2 and 3.
1861. Utilitarianism. In *Collected works*, vol. 10.
1865. Auguste Comte and positivism. In *Part I, collected works*, vol. 10.
1873. Autobiography. In *Collected works*, vol. 1.
1963–91. In *Collected works of John Stuart Mill,* ed. J.M. Robson, 33 vols. Toronto/ London: University of Toronto Press/ Routledge & Kegan Paul.

Bibliography

Bentham, J. 1827. In *Rationale of judicial evidence, specially applied to english practice,* ed. J.S. Mill, 5 vols. London: Hunt & Clarke.
De Marchi, N. 2002. Putting evidence in its place: John Mill's early struggles with 'facts in the concrete'. In *Fact and fiction in economics. Models, realism, and social construction*, ed. U. Maki. Cambridge: Cambridge University Press.
Dumont, P.É.L. 1802. *Traité de législation civile et pénale,* vol. 3. Paris: Boussange, Masson, Besson.
Hayek, F.A. 1944. *The road to serfdom*. London: Routledge.
Hollander, S. 1985. *The economics of John Stuart Mill*. Oxford: Basil Blackwell.
Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan.
Lardner, D. 1850. *Railway economy*. New York: A.M. Kelley, 1968.
Maas, H. 2005. *William Stanley Jevons and the making of modern economics*. Cambridge: Cambridge University Press.
Malthus, T.R. 1823. *The measure of value stated and illustrated*. London: John Murray.
Malthus, T.R. 1827. *Definitions in political economy*. London: John Murray.
Marshall, A. 1879. In *The pure theory of foreign trade*. *The pure theory of domestic values*. Circulated privately; repr. London: London School of Economics and Political Science, 1930.
Marshall, A. 1890. In *The principles of economics*, ed. C. Guillebaud, vol. 2. 9th (variorum) ed. London: Macmillan for the Royal Economic Society, 1961.
Mill, J. 1821. *The elements of political economy*. London: Baldwin, Cradock & Joy.

Ricardo, D. 1817. *The principles of political economy and taxation*. London: John Murray.
Robson, J.M. 1968. *The improvement of mankind*. London: Routledge & Kegan Paul.
Smith, A. 1759. In *The theory of moral sentiments*, ed. A.L. Macfie and D.D. Raphael. Oxford: Oxford University Press, 1976.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell and A.S. Skinner, vol. 2. Oxford: Oxford University Press, 1976.
Tooke, T. 1826. *Considerations on the state of the currency*. London: John Murray.
Tozer, J.E. 1838. *Mathematical investigation of the effect of machinery on the wealth of a community in which it is employed and on the fund for the payment of wages*. Cambridge: Cambridge Philosophical Society, Transaction 6.

Mill, John Stuart, as Economic Theorist

Samuel Hollander

John Stuart Mill was born on 20 May 1806 to James and Harriet (Burrow) Mill in Pentonville, London; and died on 7 May 1873 in Avignon. He was educated privately by his father on Benthamite pedagogic principles. At seventeen he joined his father at the East India Company as junior clerk, retiring as Chief Examiner in 1858. In 1824 appeared the first of many contributions to the *Westminster Review.* Mill directed the *London Review* (*London and Westminster Review* 1836) from 1834 till 1840. He sat as Member of Parliament for Westminster from 1865 to 1868.

The Ricardian Paradigm

J.S. Mill insisted on the Ricardian character of his economic theory: 'I doubt if there will be a single opinion (on pure political economy) in the book [*Principles of Political Economy* (1848)] which may not be exhibited as a corollary from his [Ricardo's] doctrines' (letter of 22 Feb. 1848; *CW*, XIII, p. 731). He did not ignore the criticisms of the preceding quarter century by 'dissenting' critics of Ricardo, but (quite correctly) did not believe them to be destructive of the main Ricardian theoretical structure (1845b, pp. 395–6; cf. Hollander 1977). From Mill's perspective, the core of the Ricardo doctrine amounted to the proposition that an increase in the general wage rate generates a fall in the general rate of profit on capital rather than an overall increase in manufacturing prices (and reduced rent in agriculture) as Adam Smith had maintained (letter of 4 Oct. 1872; *CW*, XVII, pp. 1909–10).

In Ricardo's formulation of this inverse wage-profit relation a role is played by the 'absolute standard of value' – a commodity ('gold') produced by a constant quantity of labour, and acting therefore as a labour-measuring device. An increase in the labour embodied in the wage-basket will be reflected in an increase in the gold wage and will necessarily entail an increase in the share of wages in any given value of output (output produced by a given labour input) available for distribution between labourers and capitalists. (The return to landlords is excluded by treating rent as a differential surplus and attending to the marginal product in agriculture; land is presumed not to contribute to manufacturing.) Ricardo's attention was upon *per capita* wages: an increase in *per capita* 'gold' wages implies an increase in the (proportionate) share of wages in *per capita* output which is of constant 'value' since it is the result of a specific input of labour, and a corresponding decrease in the (proportionate) share of profits. The rate of profit on capital is taken to be a direct function of the latter. The Ricardian scheme thus relates the rate of return on capital to the labour embodied in *per capita* wages – i.e. to the proportion of the work-day devoted to the production of wages, a proposition which has a strong Marxian flavour (Hollander 1979, ch. 5).

Ricardo's analysis applies whether the wage change reflects an altered wage basket due in turn to altered demand–supply conditions in the labour–market (such as, on the side of labour demand, a change in the rate of saving, or new labour-displacing technology, or an altered pattern of consumer tastes involving products produced by differential factor ratios) *given*

productivity or an unchanged (or even a falling) wage-basket with *decreasing* productivity, It will be noted that though profits appear to be a 'residual' income ('profits depend on wages'), Ricardo allowed that the profit rate acts upon the rate of savings *via* 'the motive to accumulate'. According, labour demand and the commodity wage rate are affected by alterations in the rate of profit. Profits are, therefore, a residual only in the formal sense that the sole contractual payment is that to labour, but not in the substantive sense of a 'surplus value'.

The famous application of Ricardian theory to the problem of corn-import restriction, a central policy issue, is but one of various applications of the fundamental theorem. In this particular application, which pertains to a growing (and closed) economy, the commodity wage falls as the rate of capital accumulation (and consequently demand for labour) decelerates because of land scarcity (diminishing agricultural returns), and checks the rate of population growth. But the 'money' wage rises – reflecting increased labour embodied in the smaller basket thus reducing the general return on capital. The process continues until the commodity-wage and profit rate attain their respective minima, when both population and capital cease to grow – the stationary state (Hicks and Hollander 1977).

As already intimated, the theory of value served as foundation of the analysis of distribution. More precisely: Ricardo sought to define the minimum conditions required of a medium of exchange which would assure constancy in the value of output to be shared between the income recipients in the face of a change in distribution (cf. Sraffa 1951). Only in the event of uniform capital-labour ratios in all sectors will a simple labour theory of exchange value apply, such that exchange rates are invariable to wage changes. Ricardo appreciated that in the presence of non-uniform factor proportions a wage-increase impinges differentially on costs, and thus long-run prices, depending upon the labour intensities of various sectors. He frequently proceeded (as in the above account we have proceeded) by implicitly presuming uniformity. On the other occasions he assumed a medium with mean factor proportions in which case a wage increase would cause some prices to rise and others to fall in terms of that medium, though to the extent that these variations cancel out the basic theorem remains more-or-less intact. It must at the same time be emphasized that the general conclusion whereby the rate of return is governed by the proportion of the work-day devoted to the production of wages was envisaged as holding good quite generally, even where wages and profits are expressed in terms of ordinary money, both metallic and paper. Ricardo's model was designed to throw light on the underlying processes, not always apparent in a modern capitalist-exchange economy, whereby the rate of return is governed by the proportion of the work-day devoted to the production of wage goods.

Ricardo's analysis of the determination of relative prices implies a system of economic organization directed by price forces, for he assumes the possibility of output expansion and contraction in response to market signals within a competitive framework, adopting Adam Smith's analysis of the relation between (short-run) market prices and (long-run) cost prices, the latter characterized by equality of wage-rates and of profit-rates across all sectors such that when market prices everywhere equal cost prices there is no motive for reallocation. In the case of uniform factor ratios, a wage increase generates no factor reallocation and thus no long-run price variation precisely because 'the cause that operates on one [industry] operates on all; how then can it be said that the relative values of commodities will be affected?' (1951; II, p. 179). Where factor inputs differ, labour-intensive industries will be impinged upon more than others, the relative profit rates in those sectors will fall (at the original prices) more sharply, and factors will transfer to sectors less severely affected; in consequence of these factor movements, prices rise (in terms of the measure produced with mean factor proportions) in the contracting labour-intensive sectors and fall in the expanding capital-intensive sectors, an outcome hinging upon the standard (Smithian) assumption of negatively-sloping demand curves.

To summarise: While Ricardo's major preoccupation was the 'macro-economic issue of the relation between the *general* wage-rate and the *general*

profit-rate he was obliged to deal with the structure of the economy and this problem he approached from a 'general-equilibrium' perspective. This latter perspective explains his explicit subscription to J.B. Say's account (1819) – which has Smithian pedigree – of mutual interdependence between product and factor markets incorporating both opportunity cost and the imputing of factor values from product values (cf. Ricardo 1951; I, 282). It remains to add that Say's 'Law of Markets' was used to close the Ricardian 'general-equilibrium' system (Hollander 1979; ch. 6).

Mill on Value and Distribution

We have defined what came to be known after 1817 as the 'New Political Economy' to describe Ricardo's particular contribution. In an essay 'On Profits and Interest' Mill presents favourably the Ricardian position, with the 'slight modification' that the rate of profit is related not to the value of *per capita* wages – the direct and indirect labour embodied in the wage bill – but to the 'cost of wages' which includes the profit of the wage-goods producer (1844c; pp. 293ff). But even this modification is withdrawn in Book II of the *Principles* where the profit rate is related inversely to the fraction of a man's labour time devoted to the production of his wages. The 'cost of labour' is thus finally identified with labour embodied in *per capita* wages and with labour's share in *per capita* output (1848; pp. 411ff).

This analysis of profits was provisional: 'It will come out in greater fullness and force when, having taken into consideration the theory of Value and Price, we shall be enabled to exhibit the law of profits in the concrete – in the complex entanglement of circumstances in which it actually works' (p. 415). Throughout his career, Mill insisted upon the *relativity* of exchange value and, like Samuel Bailey (1825), rejected the notion of a general alteration in exchange value as logically incomprehensible. But he accepted the Ricardian measure of cost of production:

> [Economists] have imagined a commodity invariably produced by the same quantity of labour; to which supposition it is necessary to add, that the fixed capital employed in the production must bear always the same proportion to the wages of the immediate labour, and must be always of the same durability: in short, the same capital must be advanced for the same length of time, so that the element of value which consists of profits, as well as that which consists of wages, may be unchangeable (1848; 'Of a Measure of Value' Book III, xv, p. 579).

(Missing here is the condition that the metal be produced by average factor proportions, but Mill may have been presuming uniform factor ratios.) Now such a measure of cost 'though perfectly conceivable' would not probably be found in practice because of the high likelihood of changes in the production cost of any commodity chosen. Nevertheless, gold and silver 'are the least variable' and, if used, the results obtained must simply be 'corrected by the best allowance we can make for the intermediate changes in the cost of the production itself'.

The full analysis of the effects of wage-rate changes is undertaken in the important chapter 'Distribution, as affected by Exchange' (ch. xxvi). Much is made by commentators of Mill's treatment of Production, Distribution, and Exchange in three consecutive books, as indicative of a failure to envisage any relation between value theory and distribution. This is a misunderstanding. The initial discussion of distribution in Book II was provisional only; in the chapter at hand the order is reversed and the problem of distribution is analysed in the light of the theory of exchange value.

When the distribution of national income occurs via the mechanism of exchange and money, the 'law of wages' remains unchanged insofar as the determination of commodity wages is concerned, for this depends upon 'the ratio of population and capital' (p. 695). But (as Mill has already explained) from the perspective of the employer it is not merely *commodity* wages that are relevant but the 'cost of labour'; the added point is that this cost will be reflected by the *money* wages paid when money constitutes 'an invariable standard':

> Wages in the second sense [cost of labour], we may be permitted to call, for the present, money wages; assuming, as it is allowable to do, that money

> remains for the time an invariable standard, no alteration taking place in the conditions under which the circulating medium itself is produced or obtained. If money itself undergoes no variation in cost, the money price of labour is an exact measure of the Cost of Labour, and may be made use of as a convenient symbol to express it (p. 696).

Assuming money to be such an invariable measure, the rate of money wages will depend upon the commodity wage and the production costs (and accordingly the money prices) of wage goods, particularly agricultural produce, which vary with 'the productiveness of the least fertile land, or least productive agricultural capital' (p. 697). Since the cost of labour is equated with the proportionate share of the labourer in *per capita* output, Mill had fully subscribed to the fundamental Ricardian theorem on distribution involving a 'proportions-measuring' money in terms of which a rise of wages implies an increased share of the labourer in the 'value' of his output and a reduced profit share and rate of return: 'If the labourers really get more, that is, get the produce of more labour, a smaller percentage must remain for profit. From this Law of Distribution . . . there is no escape. The mechanism of Exchange and Price may hide it from us, but is quite powerless to alter it' (pp. 479–80). The 'Marxian' flavour of this formulation may be reinforced by Mill's proposition that 'the cause of profit' can be traced to surplus labour time – the fact that labourers 'in addition to reproducing their own necessaries and instruments, have a portion of their time remaining, to work for the capitalist' (p. 411; first introduced in 4th edition of 1857). For Mill, however, as for Ricardo, the rate of accumulation (and therefore the demand for labour and the commodity wage) responds to variations in the profit (interest) rate since savers must be compensated for the psychic cost of abstaining from present consumption ('abstinence'). The breakdown between 'necessary' and 'surplus' labour time is, therefore, a variable dependent upon the supply conditions of capital as well as of labour (population).

As in Ricardo's formulation, the proposition that an increase in the labour embodied in wages is necessarily accompanied by an inverse movement in the rate of return holds good irrespective of the satisfaction by the medium of exchange of the necessary properties required to guarantee its theoretical suitability as invariable standard. Thus even were prices to rise following an increase of wages, producers would not benefit therefrom since all their expenses rise (p. 479). More significantly, the gold standard mechanism assured that wage increases are non-inflationary: 'There cannot be a general rise of prices unless there is more money expended. But the rise of wages does not cause more money to be expended' (1869, p. 661).

Mill and the Theory of Allocation

As in Ricardo's case, allocation theory provided the primary rationale for the operation of the inverse wage–profit relation. To this matter we turn next.

The theory of costs was treated by Mill, in Ricardian fashion, from a micro-economic perspective involving relative value: 'Value is a relative term, not a name for an inherent and substantive quality of the thing itself' (1848, p. 479). Accordingly, he defended Ricardo's emphasis upon labour-quantity on the grounds that 'In considering . . . the causes of *variations* in value, quantity of labour is the thing of chief importance; for when that varies, it is generally in one or a few commodities at a time, but the variations of wages (except passing fluctuations) are usually general, and have no considerable effect on value' (p. 481). Nonetheless, wage differentials as well as differential labour input are reflected in the price structure, and changes in wage differentials will generate changes in the price structure (p. 480; also p. 692). Moreover, in consequence of differential factor propositions, even general wage changes might influence the structure of prices (p. 484).

Notwithstanding Malthus's early interpretation to the contrary (1824), Mill insisted that, in the opinion of 'the Ricardo school', long-run cost prices are arrived at by way of supply variation (1825a, pp. 33–4). In the *Principles* he cautioned what while 'the value at any particular time is the

result of supply and demand, unless that value is sufficient to repay the Cost of Production, and to afford, besides, the ordinary expectation of profit, the commodity will not continue to be produced'. Necessary price, in brief, includes a return on capital 'as great ... as can be hoped for in any other occupation at that time and place'; and in the event of a return in excess of the going rate 'capital rushes to share in this extra gain, and by increasing the supply of the article, reduces its value'; in the reverse case output is restricted (1848, pp. 471–2). By this reference to 'a law of value anterior to cost of production, and more fundamental, the law of demand and supply' (p. 583), Mill did not, any more than Ricardo, deny that cost of production works its influence by way of supply variation; but maintained that demand–supply analysis applied to all cases, even where cost analysis is irrelevant.

The central role of supply variation in the establishment of cost price is scarcely surprising considering that the pertinent perspective in cost–price analysis is one involving 'the motives by which the exchange of commodities against one another is immediately determined' (letter of 15 May 1872; *CW*, XVII, p. 1895).

Following Ricardo, Mill employed this perspective in the rationalization of the inverse wage–profit relation. In contrast to a wage increase affecting one sector where price will rise to assure equality of profit rates across the board (or a general wage increase in the case of non-uniform factor proportions), there exists no allocative mechanism whereby general prices would be forced upwards in the event of an economy-wide wage increase, should all firms be affected equally by the change: 'There is no mode in which capitalists can compensate themselves for a high cost of labour, through any action on values or prices. It cannot be prevented from taking its effect in low profits' (1848, p. 479).

The Ricardo–Mill allocation mechanism implies negatively sloped market demand curves. Mill took this for granted: 'It is the next thing to impossible that more of the commodity should not be asked for at every reduction of price' (1869; p. 637). Mill's formulations constituted an improvement in rigour over Ricardo's – particularly the formal conception of an equation of demand and supply and the distinction between displacements of the demand schedule and movements along the same schedule (1848, p. 466). But their merit reflects less innovatory content than location at a conspicuous juncture amongst the basic theoretical principles. There are brilliant applications of demand–supply analysis to the joint-production case (pp. 582f), and to international trade (1844a; 1848, pp. 587f) – specification of the terms of trade emerging between the limits imposed by the autarkic cost ratios established by Ricardo and the division of the grains from trade, constrained only by a failure to fulfill a promise to show how the range of indeterminateness can be removed in cases of multiple or neutral equilibrium.

The analysis of rent provides a further instance of Mill's elaborations regarding allocation theory. In the aggregate, rent differs from the other factor returns solely in consequence of given land supply (1848, p. 58). Allowing for qualitative differentials between plots complicated the issue only slightly (p. 429); Mill, following Ricardo, realized that differential rent entails a special case of scarcity value, and that rent might be generated even in the absence of differentials in the event of an absolute constraint on farm output (p. 428). But when he focused upon individual sectors, he spelled out (as Smith and Say had done but Ricardo had failed to do) the consequence of multi-use land for cost pricing (p. 498; cf. p. 494), although this perspective plays no part in the analysis of wage and profit rates and their secular movement, where rent is treated entirely as a differential surplus.

Consistent with Mill's 'Ricardian' approach to cost price (exception made for the multi-use land case) is the Smith-Say conception of organization which emphasizes the ultimate source of factor remuneration in sales proceeds and the motive for factor employment in the revenue product; 'in the present system of industrial life, in which employments are minutely subdivided ... all concerned in production depend for their remuneration on the price of a particular commodity' (1848, p. 455); transportation workers 'derive their remuneration from the ultimate product'

(p. 33); in consequence of the 'increased utility' afforded by wholesales and retailers the product is sold 'at an increased price proportioned to the labour expended in conferring it' (p. 48).

The principle that the process of production ends upon sale to the final consumer applies also to wage goods (pp. 35, 38). While 'the finished products of many branches of industry are the materials of others' (p. 36), workers' consumables are treated on a par with all other final goods rather than as intermediate goods. In Mill's account (as in Ricardo's) workers are paid in money, not in kind, and enter the market to purchase commodities at retail; there is no distinction to this regard between labourers, capitalists or landlords. The 'wages fund' expressed in money has a real counterpart in the flow of goods currently made available at retail outlets; the fraction of capital whose function it is to fulfill the tasks of 'maintaining' labour need not actually take the form of stocks of wage goods, because the flexibility of the system permits, by exchange or production, the easy and rapid generation of commodities in demand by labour (pp. 57, 67–8, 82–3).

The principle of imputation applies to the demand for particular kinds of labour or labour in particular industries. By contrast, Mill's proposition that 'demand for commodities is not demand for labour' (1848, p. 78) relates to *aggregate* wages and/or employment: 'it is only by what [a person] abstains from consuming, and expends in direct payments to labourers in exchange for labour, that he benefits the labouring classes, or adds anything to the amount of their employment' (p. 80). Both Ricardo and J.B. Say were said to have fully appreciated this position. It is to be noted that when capitalist employers make an investment decision they abstain from using their own claim to purchase output currently forthcoming at retail outlets and place this purchasing power at the disposal of labourers (pp. 83–4).

Mill on Growth, the Cycle and the Law of Markets

In his approach to growth, Mill, following Malthus, supplemented the Ricardian analysis involving a simultaneous decline in both the real wage and the profit rate until their respective minima in circumstances of land scarcity. Mill demonstrated that in a situation of growing capital and population the commodity wage need not decline to 'subsistence' if labourers respond to the prospective decline in the rate of accumulation by delaying marriage and reducing procreation. A fall in the wage rate is then no longer necessary to reduce population growth in line with the rate of accumulation. The fall in profits will, however, be more rapid and the stationary state achieved sooner than in the Ricardian version. This model (cf. Hollander 1984, 1985a, pp. 444–51) provides the theoretical backdrop to Mill's reconsideration of the possible merits to zero growth.

The idea of an endogenous trade cycle turning on expectational mood is better developed by Mill than any contemporary. The regularity of cyclical fluctuations was much emphasized in a monetary paper of 1844 (1844d). In the *Principles* Mill attended to the 'quiescent' period and its place in the cycle. Specifically, a quiescent period entails expansion rather than stationarity, and cyclical fluctuations are partly induced by speculative reactions to the falling return on capital arising from 'the gradual process of accumulation' (1848, p. 641). The relationship is a mutual one, for while the declining profit-rate trend engenders speculation and the cycle, various losses associated with the cycle play back on the profit rate itself.

In the absence of capital loss the rate of accumulation would be so great (on Mill's empirical estimate) as to force down the return on capital since technical progress could not in practice be relied upon to counteract such heavy pressure on scarce land. The first conclusion Mill draws from the fact of a highly active contemporary 'spirit of accumulation' is that 'a sudden abstraction of capital, unless of inordinate amount', need not be feared, for 'after a few months or years, there would exist in the country just as much capital as if none had been taken away' (p. 747). The conclusion altered the perspective towards government expenditure. The standard warnings by orthodox writers against measures which might reduce the capital stock, or its rate of accumulation, were no longer pertinent. Indeed, Mill writes

in this context as if capital is no longer to be treated as a scarce factor.

The question arises whether Mill's favourable attitude towards expenditure of public money 'for really valuable, even though industrially unproductive purposes', has genuine Keynesian overtones. The answer must be in the negative. The potential problem is excessive accumulation forcing down the return on capital in the Ricardian fashion – excessive in the sense that the pressure on land exceeds the counteracting force of new technology. Such a decline in the return is in practice temporary, however, in consequence of capital losses – poorly considered 'speculative' additions to the real capital stock which prove untenable in quiescent periods (the speculations induced to some degree by the temporary fall in the profit rate) and the running down of savings for consumption purposes in depression, the inevitable sequel to speculative periods. To this extent there is no question of leakages from the income stream by the non-investment of savings; savings are lost in the sense only of being unproductively used up. Mill's allowance for higher government spending thus amounts to a recommendation to tap the flow of savings, thereby preventing their excessive accumulation, pressure on scarce land and fall in the return on capital and also the various cyclical consequences of that fall which include wastage of capital. In effect, Mill was calling for opera houses in place of a superfluous network of railways and 'unproductive' private consumption. This is not a 'Keynesian' perspective.

The orthodox law of markets is in one sense firmly reiterated: there can be no 'overproduction'. But excess capacity and excess supplies of labour and commodities with a counterpart in an excess demand for money to hold are fully allowed as a feature of depression (1844b), a remarkable case of model improvement. At the same time Mill explained why stagnation would be temporary, by reference to a reversal of expectations which encourages a delay of sales wherever possible and a renewal of purchases in response to prospective price increases. This is the basis for Mill's presumption against a Keynes-like 'unemployment equilibrium,' and explains partly why government expenditure was not envisaged as a counter-cyclical measure. Only indirectly would government spending by effective, for by imposing a floor to the return on capital it checks the 'speculative fever' from which depression ultimately proceeds.

It has been well said that Mill's qualifications to the law of markets lead one 'to wonder why so much of the subsequent literature . . . had to be written at all' (Baumol and Becker 1952). The recognition of excess demand for money extends to an allowance for active monetary policy to mitigate cyclical pressure. It is regrettable that later economists felt able to brush aside the classical contribution. Mill's warning against overfull employment and his denial of a permanent trade-off between inflation and unemployment (1833) also bear repeating in our day.

Concluding Note

John Stuart Mill's methodological perspective (1836) took a stand against professional arrogance and narrow-mindedness. He justified a specialist economics on empirical grounds, and disdained all notion of the universal validity of axioms. He invited consideration of the functioning of an economic system under a variety of alternative institutional arrangements and alternative circumstances, including the 'stationary state', although his concern for equitable distribution did not lead him to dispose of the old growth economics. He maintained a modest estimate of the predictive potential of economic science. He recommended model improvement by way of verification against factual evidence, and focused on the mechanics of pricing in the real world of business rather than some ideal world. He feared the kind of applied mathematical research programme already under way during his last years.

As he, and later Marshall, always insisted Mill on pure theory (as well as on method) was Ricardian. The analytics of Marshall's *Principles* are in a 'direct line of descent through Mill from Ricardo' (Shove 1942). This generalization applies preeminently to the theory of value and distribution, for classical cost-price analysis constitutes an analysis of the allocation of scarce

resources, with allowance for final demand and the interdependence of factor and commodity markets. Mill's contribution to international trade theory is but an outstanding instance of a broad comprehension of demand theory. The demand-oriented economists of the 1870s exaggerated the innovatory character of their contributions. Similarly Mills supply and demand determination of wages and profit is in a line common to Ricardo (and before him Smith), and Marshall.

Mill's perspective on growth – his allowance for progress to the stationary state without depression of the real wage reflects the perspective of the Philosophical Radicals and, before them, Malthus himself, on desirable social policy. This issue illustrates well the character of classical theory as an exercise in persuasion designed to act on key behavioural patterns rather than as a 'predictive' device; theory suggested not what *will* happen, but what, depending on circumstances, *can* happen (Shackle, 1972, pp. 72–3).

See Also

- ▶ British Classical Economics
- ▶ International Trade
- ▶ Marshall, Alfred (1842–1924)
- ▶ Offer Curve or Reciprocal Demand Curve
- ▶ Ricardo, David (1772–1823)

Selected Works

CW = Collected Works of John Stuart Mill, ed. J.M. Robson, Toronto: University of Toronto Press, 1963–85.

1824. War expenditure. *Westminster Review* 2: 27–48. In *CW*, IV, 1967, 1–22.

1825a. The Quarterly review on political economy. *Westminster Review* 3: 213–232. In *CW*, IV, 1967, 23–43.

1825b. The Corn laws. *Westminster Review* 3: 394–420. In *CW*, IV, 1967, 45–70.

1826. Paper currency and commercial distress. *Parliamentary Review Session of 1826,* 630–662. In *CW*, IV, 1967, 71–123.

1833. The currency juggle. *Tait's Edinburgh Magazine* 2: 461–7. In *CW*, IV, 1967, 181–92.

1834. Miss Martineau's summary of political economy. *Monthly Repository* 8: 318–322. In *CW,* IV, 1967, 223–228.

1836. On the definition of political economy; and on the method of philosophical investigation in that science. *London and Westminster Review* 4(26): 1–29. (Appears as Essay V in *Essay on Some Unsettled Questions of Political Economy.* 1844, with title . . . and on the method of investigation proper to it.) In *CW*, IV, 1967, 309–339.

1844a. Of the laws of interchange between nations. In *Essays on some unsettled questions in political economy*, ed. J.S. Mill. London: Parker. In *CW*, IV, 1967, 232–261.

1844b. Of the influence of consumption on production. In *CW*, IV, 1967, 262–279.

1844c. On profits and interest. In *CW*, IV, 1967, 290–308.

1844d. The currency question. *Westminster Review* 41: 579–98. In *CW*, IV, 1967, 341–361.

1845a. The claims of labour. *Edinburgh Review* 81: 498–525. In *CW*, IV, 1967, 363–389.

1845b. De Quincey's logic of political economy. *Westminster Review* 43: 319–331. In *CW,* IV, 1967, 391–404.

1848. Principles of political economy with some of their applications to social philosophy. In CW, II-III, 1965. Last (7th) edn by Mill, 1871.

1869. Thornton on labour and its claims. *Fortnightly Review,* New Series 5, May, 505–18; June, 680–700. In *CW*, V, 1967, 631–68.

1879. (Posthumous.) Chapters on socialism. *Fortnightly Review,* New Series 25, February, 217–37; March, 373–82; April, 513–30. In *CW*, V, 1967, 703–53.

Bibliography

Bailey, S. 1825. *A critical dissertation on the nature, measure and causes of value*. London: R. Hunter.

Baumol, W.J., and G.S. Becker. 1952. The classical monetary theory: The outcome of the discussion. *Economica* 19(76): 355–376.

Hicks, J.R., and S. Hollander. 1977. Mr. Ricardo and the moderns. *Quarterly Journal of Economics* 91(3): 351–369.

Hollander, S. 1977. The reception of Ricardian economics. *Oxford Economic Papers* 29(2): 221–257.
Hollander, S. 1979. *The economics of David Ricardo*. Toronto: University of Toronto Press.
Hollander, S. 1984. The wage path in classical growth models: Ricardo, Malthus and Mill. *Oxford Economic Papers* 36(2): 200–212.
Hollander, S. 1985a. *The economics of John Stuart Mill*. Oxford: Basil Blackwell.
Hollander, S. 1985b. On the substantive identity of the Ricardian and neo-classical conceptions of economic organization. In *The legacy of Ricardo*, ed. G. Caravale. Oxford: Basil Blackwell.
Malthus, T.R. 1824. Political economy. *Quarterly Review* 30: 297–334.
Ricardo, D. 1951. In *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.
Say, J.B. 1819. *Traité d'économie politique.* 4th ed. Paris.
Shackle, G.L.S. 1972. *Epistemics and economics: A critique of economic doctrines*. Cambridge: Cambridge University Press.
Shove, G. 1942. The place of Marshall's *Principles* in the development of economic theory. *Economic Journal* 52: 294–329.
Sraffa, P. 1951. Introduction. In *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.

Millar, John (1735–1801)

Nicholas Phillipson

Born in Lanarkshire, the son of a Presbyterian minister, Millar was educated at Glasgow University for the Scottish Bar. He became a protégé of Adam Smith and Lord Kames, both of whom were instrumental in securing his appointment as Professor of Civil Law at Glasgow, a post he held until his death. He was a charismatic teacher who transformed the civil law curriculum by placing it on the jurisprudential foundations Smith had created for his moral philosophy lectures. He was a radical Foxite Whig and a member of the Society of the Friends of the People.

Millar is now much admired by historians of social thought for his *Origin of the Distinction of Ranks* (1771) in which he appeared to develop the sociological implications of Smith's account of the progress of civilization in a history of different systems of social authority. Unfortunately this view will not stand. The publication of the text of Smith's *Lectures on Jurisprudence* in 1978 showed that Millar's apparently original analysis, for all its closeness of texture and acuity, was intellectually entirely dependent on Smith's earlier work.

He never gave a separate course of lectures on political economy, and dealt with that subject in unpublished lectures on government whose character can be inferred from a series of essays first published in the posthumous edition of his *Historical View of the English Government* (1803). His Smithian interest in the natural history of property led him to formulate a distinctive theory of profit as the wage of the manufacturer plus the saving he derived from investing in the division of labour, a subject which also interested his pupil, the Earl of Lauderdale. The attraction of this theory lay in its radical political implications. It allowed Millar to show that the attributes of property ownership and personal independence which lay at the heart of contemporary ideas of political rights extended to all of those who participated in the productive relationships of a commercial society. It led him to campaign for the radical reform of parliament in order to adjust the old Whig constitution to the social and economic changes of the past century.

See Also

► Enlightenment, Scottish

Selected Works

1771. *The origin of the distinction of ranks*, 4th ed. Edinburgh: William Blackwood; London: Longman, Hurst, Rees & Orme, 1806.
1803. *An historical view of the English government*, 4 vols. London: J. Mawman.

Bibliography

Igantieff, M. 1983. John Millar and individualism. In *Wealth and virtue: The shaping of political*

economy in the Scottish enlightenment, ed. I. Hont and M. Ignatieff. Cambridge: Cambridge University Press.
Lehmann, W.C. 1960. *John Millar of Glasgow. 1735–1801: His life and thought and his contributions to sociological analysis*. Cambridge: Cambridge University Press.

Miller, Merton (1923–2000)

René M. Stulz

Abstract

Merton Miller was at the centre of the transformation of academic finance from a descriptive field to a science. His principal contribution to this transformation was the introduction of arbitrage arguments which underlie most theoretical contributions in finance and remain central to the way financial economists analyse finance problems to this day. These arbitrage arguments underlie his and Franco Modigliani's famous irrelevance propositions.

Keywords

American Finance Association; Arbitrage; Bankruptcy costs; Capital gains taxation; Capital structure; Contracting costs; Corporate bonds; Corporate debt; Corporate finance; Dividend policy; Financial markets; Interest rates; Rrrelevance propositions; Miller, M.; MM irrelevance propositions; Modigliani, F.; Risk; Tax subsidy; Tax shield; Taxation of corporate profits

JEL Classifications

B31

From the late 1950s to the early 1970s the field of finance changed fundamentally. A reader of the *Journal of Finance* in the early 1950s would find a field that was mostly descriptive. After the early 1970s the field had become a science. Merton Miller was at the centre of that transformation. His work started it in 1958. For the rest of his life he was at the heart of modern finance. (Grundy 2001, provides a complete list of Merton Miller's publications.)

After obtaining a Ph.D. in economics from Johns Hopkins University in 1952 and a brief stay at the London School of Economics, he joined the Graduate School of Industrial Administration at what was then known as Carnegie Tech. As an assistant and associate professor there, he made the contributions to the theory of corporate finance with Franco Modigliani, another faculty member, that made him famous. He joined the University of Chicago in 1961. From Chicago he exerted a huge influence on finance which lasted until he died in 2000. Merton Miller's research had a prodigious impact – he made major contributions in monetary economics, operations research, derivatives pricing, and asset pricing, as well as his seminal contributions in corporate finance – but his influence went far beyond the contributions of his papers. He mentored countless Chicago graduate students and faculty members from Chicago and throughout the profession. At times he played the role of the nurturing patriarch, while at other times he used his wit and intellect to keep people on the straight and narrow path of solid economic thinking. From 'his' seat on the left of the speaker in the Rosenwald seminar room, often in a worn-out sweater, he changed the course of numerous papers. Sometimes his intervention went further – for example, he was instrumental in persuading the *Journal of Political Economy* to publish the paper by Black and Scholes that is the foundation of option pricing theory. When he ventured outside of the University of Chicago, he often did so to be 'an activist supporter of free-market solutions to economic problems', as he stated in a brief Nobel autobiography (1991a). He knew how to make his case – he was not the son of an attorney, a Harvard graduate also, for nothing – and had a well-deserved reputation for unparalleled eloquence in the finance profession.

The Irrelevance Propositions and the Role of Arbitrage

Merton Miller earned a Nobel Prize in economics in 1990 for his 'fundamental contributions to the theory of corporate finance' (Franco Modigliani already had a Nobel Prize by then for his life cycle theory of saving). Just about every MBA in the world has learned the famous MM irrelevance propositions he developed with Franco Modigliani. (One paper had Modigliani's name first and the other had Miller's name first, so I will proceed using the moniker MM to represent the team.) The two key MM irrelevance propositions are developed in a world with perfect markets, so that there are no frictions. In particular, there are no transactions costs or taxes, and no costs are incurred to induce managers to maximize the value of the firm.

The first irrelevance proposition, Proposition I in the paper titled 'The Cost of Capital, Corporation Finance and the Theory of Investment' published in the *American Economic Review* (1958, p. 268) states that 'the market value of any firm is independent of its capital structure and is given by capitalizing its expected return at the rate. ... appropriate to its class'. The second irrelevance proposition concludes that 'given a firm's investment policy, the dividend payout it chooses to follow will affect neither the current price of its shares nor the total return to its shareholders' (1961, p. 414). In other words, in perfect markets neither capital structure choices nor dividend policy decisions matter. Since then, corporate finance has refined these results and built theories based on the existence of market imperfections.

If we had to remember one thing about Merton Miller's contributions to finance, what should it be? It would not be the irrelevance propositions themselves. Rather, it would be the way the irrelevance propositions were proved (for a more complete analysis, see Stulz 2000). The approach used to prove these propositions is central to the thinking of practitioners of modern finance. It has spawned many seminal contributions to the field. The method used to prove Proposition I is the method of arbitrage. MM did not invent arbitrage, but made it the foundation of modern finance. MM assume that financial markets are perfect and then show that

> if Proposition I did not hold, an investor could buy and sell stocks and bonds in such a way as to exchange one income stream for another stream, identical in all relevant respects but selling at a lower price. The exchange would therefore be advantageous to the investor quite independently of his attitudes toward risk. As investors exploit these arbitrage opportunities, the value of the overpriced shares will fall and that of the underpriced shares will rise, thereby tending to eliminate the discrepancy between the market values of the firms. (1958, p. 269)

The arbitrage mechanism is how Merton Miller thought about finance phenomena. Results that would lead to arbitrage opportunities could not possibly be important because market forces would step in to make prices right. However, in his thinking arbitrage was never limited to existing financial instruments and institutions. For him, arbitrage opportunities that exist in the real world will eventually disappear because, when needed, financial innovations will occur that will prevent these opportunities from persisting.

Though arbitrage arguments are now pervasive throughout finance and, more generally, economics, the more immediate and direct impact of the arbitrage proof of Proposition I was to provide the foundation for modern corporate finance because it specifies sufficient conditions for leverage not to matter. Because of the proof, we know that, if financial markets are perfect, the value of a firm does not depend on its leverage. As a result, practitioners and academics alike know that, if leverage affects value, it must be that one or more of the assumptions required by the arbitrage proof do not hold.

In their papers MM eliminated once and for all the argument that leverage is costly simply because it increases the interest rate the corporation pays for its debt. As leverage increases in a world of perfect markets, the coupon paid on debt increases, but that is because bondholders bear more risk and must be compensated for this additional risk. This will happen even though the firm's cash flows are unaffected by the additional leverage. Hence, as Merton Miller pointed out in

his Nobel lecture (Miller 1991c), the increase in the risk of debt has no social costs because the firm's total risk is unaffected by the change in leverage.

Beyond the Rrrelevance Propositions

With corporate income taxes, the cost of debt for the firm is the cost after taxes since interest paid on debt is tax deductible at the corporate level. If the only departure from the assumptions leading to Proposition I were a tax subsidy to corporate debt, one would expect firms to maximize the value of that subsidy and therefore have extremely high leverage. Empirically, however, leverage is not extreme. To make sense of the limited levels of leverage in the presence of what appeared to be a large tax subsidy for debt, finance had either to relax other assumptions leading to Proposition I or to conclude that the subsidy was illusory. Initially, the route chosen by finance was to take into account bankruptcy costs. Bankruptcy costs occur because contracting is costly – firms that default on their debt contracts cannot be costlessly reorganized. In the presence of bankruptcy costs and tax subsidy to debt, each firm has an optimal debt level such that the increase in the present value of expected bankruptcy costs resulting from an additional dollar of debt equals the present value of the expected tax subsidy from that additional dollar of debt.

Merton Miller always doubted that expected bankruptcy costs could be large enough to explain why firms did not take greater advantage of the tax subsidy of debt. His assessment of the evidence on bankruptcy and financial distress costs was that 'neither empirical research nor simple common sense could convincingly sustain these presumed costs of bankruptcy as a sufficient, or even as a major, reason for the failure of so many large, well-managed US corporations to pick up what seemed to be billions upon billions of dollars of potential tax subsidies' (1991b, p. 274). This assessment led him to one of his most memorable statements, namely, that 'the supposed trade-off between tax gains and bankruptcy costs looks suspiciously like the recipe for the fabled horse-and-rabbit stew – one horse and one rabbit' (1976, p. 264).

Since direct bankruptcy costs could not explain why firms were not taking advantage of the apparent tax subsidy of debt, the field of finance turned to other explanations for low leverage based on contracting costs. Jensen and Meckling (1976) showed that, as leverage increases, shareholders have incentives to take advantage of bondholders by undertaking highly risky projects with high payoffs to shareholders in some states even though such projects have a negative net present value. The bondholder–shareholder conflict identified by Jensen and Meckling makes debt more costly because firms either behave inefficiently as a result of leverage or spend real resources to convince bondholders that they will not take advantage of them. A large literature emphasizing contracting costs has developed over time.

Merton Miller always had doubts that the bondholder–shareholder conflict could explain why firms did not take greater advantage of the tax shield of debt. Not surprisingly, his scepticism stemmed from the role of arbitrage in his thinking. If the tax shield of debt was so large, why was it that investment bankers would not devise solutions that would enable firms to take advantage of this tax shield and overcome the agency costs of debt through clever contracting? As always, he viewed no finance problem as solved unless he could find a solution that would not provide clever arbitrageurs with profit opportunities.

In 1976, in his address as President of the American Finance Association, Merton Miller revisited the issue of the impact of corporate taxation on the MM irrelevance propositions in a classic paper titled 'Debt and Taxes'. This paper shows perhaps better than any of his other papers how he could use arbitrage arguments to change the way finance academics and practitioners understood how the world works. In that paper he pointed out that the tax advantage of corporate debt might be mostly if not completely illusory. Because interest on corporate debt is taxed as income to the bondholder, the interest paid must be sufficiently high to ensure that the after-tax income from holding corporate bonds is attractive relative to the income from equity which, when it

accrues as capital gains, is taxed at a lower effective rate. While corporate interest payments generate tax deductions, personal taxes on interest income are higher than on capital gains, and so the before-tax cost of capital on debt must be higher than on equity to induce investors to hold debt. In his paper Merton Miller showed that under specific conditions the only feasible equilibrium is the one in which the after-tax cost of debt equals the after-tax cost of equity. When this equilibrium obtains, Proposition I holds in the presence of taxes, and no firm has a financial incentive to alter its mix of debt and equity even though interest payments on debt are tax deductible. 'Debt and Taxes' demonstrated that the perfect-markets assumptions are sufficient, but not necessary, conditions for leverage to be irrelevant. Showing that the assumptions required for Proposition I do not hold is not enough to conclude that leverage matters; rather it must also be the case that clever arbitrageurs cannot profit from the situation.

The Legacy

With the contributions to the field of finance that I have described, Merton Miller provided a way to think about financial phenomena that remains at the core of all major theoretical developments in the field. Throughout his life, Merton Miller used arbitrage reasoning to organize his thoughts about important phenomena. His first publication appeared in the 1948 *American Economic Review*. In 1990, he published a paper in the *Journal of Finance* (co-authored with David Hsieh) that analysed the impact on stock prices of changes in margin requirements. That paper was awarded a Smith–Breeden prize for best paper in the *Journal of Finance*. At the time, Merton Miller was thrilled because he had published refereed papers in top journals in five different decades. He never stopped wanting to write papers that merited publication in top journals. Three days before his death he was preparing a paper for submission. Throughout his life he was first, last, and foremost a scholar.

See Also

- ▶ Arbitrage
- ▶ Modigliani, Franco (1918–2003)
- ▶ Modigliani–Miller Theorem

Selected works

1948. (With R. Musgrave.) Built-in flexibility. *American Economic Review* 38, 122–128.
1958. (With F. Modigliani.) The cost of capital, corporation finance and the theory of investment. *American Economic Review* 48, 261–297.
1961. (With F. Modigliani.) Dividend policy, growth, and the valuation of shares. *Journal of Business* 34, 411–433.
1976. Debt and taxes. *Journal of Finance* 32, 261–275.
1990. (With D. Hsieh.) Margin regulation and stock market volatility. *Journal of Finance* 45, 3–29.
1991a. Autobiography. In *Les Prix Nobel. The Nobel Prizes 1990*, ed. Tore Frängsmyr. Stockholm: Nobel Foundation. Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/1990/miller-autobio.html. Accessed 28 June 2006.
1991b. *Financial Innovations and Market Volatility*. Cambridge, MA/Oxford, UK: Blackwell.
1991c. Leverage. *Journal of Finance* 46, 479–488.

Acknowledgment I am grateful for comments from Harry DeAngelo, Linda DeAngelo, Steven Durlauf and Andrew Karolyi.

Bibliography

Grundy, B. 2001. M. H. Miller: His contributions to financial economics. *Journal of Finance* 56: 1183–1206.
Jensen, M., and W. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.
Stulz, R. 2000. Merton Miller and modern finance. *Financial Management* 29: 119–131.

Minard, Charles Joseph (1781–1870)

R. F. Hébert

Keywords

Cost–benefit analysis; Demand theory; Dupuit, A.-J.; Income distribution; Jevons, W.; Minard, C.; Navier, L.; Public works; Say, J.-B.; Subjective measures of utility; Substitution effect; Value of time

JEL Classifications

B31

A French engineer and economist, Charles Joseph Minard was widely recognized as the creator of graphical statistics, a means of figuratively portraying railway traffic routes on illustrated maps. Minard served as professor at the École Nationale des Ponts et Chaussées (ENPC) in the 1830s, where he taught the course on interior navigation, which included roads, rivers, canals and railways. In 1831 Minard wrote a lengthy monograph designed to establish a course in economics that he proposed for ENPC students. Although Minard viewed this work as a manual for practising engineers, J. B. Say immediately recognized the manuscript as a systematic treatise on the economics of public works, and urged Minard to publish it for the benefit of economists as well as engineers. For reasons that are not entirely clear, Minard shelved his manuscript instead – probably owing to the delay by ENPC in establishing an economics chair until 1847. In 1850, a year before his retirement from public service, Minard published his 'Notions élémentaires d'économie politique appliqué aux travaux publics' in the *Annales des Ponts et Chaussées.*

In this monograph Minard explored such fundamental notions as utility, demand, opportunity costs, the value of time and services, the effects of taxes on income distribution, and the use of compound interest in calculating the value of capital expenditures – a treatment lauded by W. S. Jevons in his *Theory of Political Economy* (1871). Despite its unfortunate delay in publication, the ideas in Minard's monograph were clearly part of the oral tradition in economics at ENPC in the first half of the 19th century. Thus, Minard served as an important link between Navier and Dupuit in the development of demand theory and cost–benefit analysis. This claim is based on four major aspects of his work: he introduced subjective elements, such as the value of time, into the operational measure of utility; he insisted that the magnitude of social utility depends on the distribution of income; he recognized that price increases cause substitution effects among existing consumers and that price decreases draw new consumers into the market; and he developed subjective notions of cost associated with public works.

See Also

- ▶ Cost–Benefit Analysis
- ▶ Demand Theory
- ▶ Dupuit, Arsene-Jules-Emile Juvenal (1804–1866)
- ▶ Jevons, William Stanley (1835–1882)
- ▶ Navier, Louis Marie Henri (1785–1836)
- ▶ Public Works
- ▶ Say, Jean-Baptiste (1767–1832)

Selected Works

1850. Notions élémentaires d'économie politique appliqué aux travaux publics. *Annales des Ponts et Chaussées: Mémoires et Documents,* 2d ser. 19(1): 1–125.

1851. Motifs pour préférer dans les travaux publics des ouvrages moins coûteux, quoique moins durables. *Journal des Economistes* 21: 65–67.

Bibliography

Coronio, G. 1997. *250 ans de L'École des Ponts en cent portraits*, 84–85. Paris: Presses de l'école des Ponts et Chaussées.

Ekelund, R. Jr., and R. Hébert. 1978. French engineers, welfare economics, and public finance in the nineteenth century. *History of Political Economy* 10: 636–668.
Ekelund, R. Jr., and R. Hébert. 1999. *Secret origins of modern microeconomics: Dupuit and the engineers*. Chicago: University of Chicago Press.
Etner, F. 1987. *Histoire du calcul économique en France*. Paris: Economica.
Hébert, R. 1994. Fondements et développements de l'économie publique. *Revue Du Dix-Huitième Siècle* 26: 37–49.
Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan.

Mincer, Jacob (1922–2006)

Finis Welch

Abstract

Jacob Mincer was one of the founding fathers of modern labour economics. Along with Gary Becker and T.W. Schultz, Mincer's ideas led to the evolution of labour economics as perhaps the premier applied field in economics. His work on personal income distributions and the associated wage–age profiles has dominated empirical research on these topics since the mid-1960s. The work extended to many related areas, most importantly the labour force participation of married women, the wage–age profiles associated with interrupted work careers, and migration decisions of two-career families.

Keywords

Age–wage profiles; Becker, G.; Labour economics; Mincer, J.; On the job training; Returns to schooling; Schultz, T. W.; Women's work and wages

JEL Classifications

B31

Born in Poland, Jacob Mincer was a college freshman in Czechoslovakia when the Germans invaded in early 1939. He spent most of the Second World War in prisons and concentration camps, but survived to enter Emory University in 1948 on an Hillel Foundation scholarship. After completing his first degree in two years, Mincer began his graduate studies in economics at the University of Chicago. He then transferred to Columbia University, having followed the lady who would be his wife from Chicago to New York for her residency in radiation oncology. Later, Flora Kaplan Mincer, MD, took six years from her practice to bring up their three children. That interruption in her career would be the basis for Mincer's subsequent paper, with his student Solomon Polachek (1974), which was the first to empirically tackle the complications of women's careers in earnings determination.

Jacob Mincer received his Ph.D. from Columbia in 1957, taught for two years at the City College of New York, and then returned to Columbia, where he remained until his 1991 retirement. In the interim, there were visiting appointments at the University of Chicago, the Stockholm School of Economics and the Hebrew University of Jerusalem.

Mincer was one of the very best 20th-century economists. He is one of the four or five who led the way into modern labour economics. The ideas of investing in man per se that were circulating in the early to mid-1950s had become environmental at Chicago and Columbia by the end of the decade. Given publication lags, it is impossible to know who came first, but the three papers that introduced the economics world to human capital were Theodore W. Schultz's 'Capital Formation by Education' (1960), Jacob Mincer's 'Investment in Human Capital and Personal Income Distribution' (1958) and Gary S. Becker's 'Underinvestment in College Education?' (1960).

Schultz argued simply that skills are malleable, that they are durable and acquired at a cost. As such they fit the capital formation rubric nicely. He also demonstrated that the opportunity costs of students who forgo work to remain in school in the aggregate are roughly equal to the costs of all purchased resources of schools and colleges. Soon afterwards he suggested that an extraordinarily large part of US per capita income growth in the first half of the 20th century was due to growth in education of the citizenry (1961a).

Mincer's 1958 *Journal of Political Economy (JPE)* paper was an extension of his thesis, which relied on the 1940, and 1950 decennial censuses. In this paper he challenged the traditional literature regarding income distributions that had focused only on the aggregate shape, with differences among individuals presumably owing only to luck and ability. After presenting a simple theory showing that with the discounted value of lifetime incomes constant, there would nonetheless be differences in income attributable to the time spent in both formal training and informal on-the-job training. The empirical support followed. In addition to laying the groundwork for what would become possibly *the* major area of empirical study in all of economics, he made the fundamental argument that the distribution of lifetime incomes is more, much more, equal than the point in time distributions. Becker simply showed that as a pure and simple investment for subsequent income alone, rates of return to a college education compare favourably to investments in physical capital. The traditional assumptions regarding the consumption and external benefits of education – for example, the ability to enjoy the arts, and so on, improved choices regarding health and life style and the externality of an informed citizenry – were not to be ignored: education is undoubtedly consumption in part, but there is also real productive value.

The introduction of the ideas of human capital to economics cumulated in T.W. Schultz's 1960 Presidential Address to the American Economics Association, 'Investment in Human Capital' (1961b), followed by the collection of articles in the October 1962 *JPE* supplement headlined with Becker's, 'Investment in Human Capital: A Theoretical Analysis' and Mincer's, 'On the Job Training: Costs, Returns, and Some Implications'. The ideas presented in these and a few of the others in the supplement, for example, Stigler's article on search, triggered an intellectual excitement and enthusiasm that coloured almost all of labour economics for the two succeeding decades. During that period, labour economics became one of the foremost applied fields of economics.

Although Mincer matched his contemporaries in insight and imagination, his work is distinguished by his insistence on empirical applicability. An elegant case in point is the piece in the 1962 *JPE* supplement. Noting that age–wage profiles have a consistent tendency to rise rapidly early in a career, less rapidly thereafter, and then stabilize or decline slightly, he characterized the shape of the profile as the result of investment in learning on the job. He began by assuming that the individual has as an option a relatively flat profile equal in discounted value to the value of the observed profile. It follows that immediately after leaving school the difference between the flat profile and the lower actual one is an investment in higher subsequent higher wages. As such, the second period opportunity wage exceeds the flat alternative by the return on the first period's investment, and so on. Assuming, further, that investment declines linearly during the early career, Mincer observed that if the rate of return on the investment in training is approximately equal to the rate used for calculating the flat alternative, then the two will intersect a number of years after leaving school, that is, approximately equal to the inverse of the rate of return. This is his famous 'overtaking point'. Some invest heavily and have a steeply inclined wage profile, while others invest less and have a profile that increases less rapidly. Even so, if the rates of return are independent of the intensity of investment, the alternative paths will intersect at the overtaking point. These are simple descriptives that are easy for graduate students to follow. Maybe that is part of the reason Mincer is so revered. The ultimate pedagogic piece came in his 1974 book *Schooling, Experience and Earnings.*

In that most influential work, Mincer specified the details of 'the' human capital earnings equation. In it the left-hand variable is the logarithm of a rate of wages or earnings. The right-hand side has years of schooling (linearly) and a quadratic is years of work experience approximated by the number of years since leaving school. In the mid-1980s Kevin Murphy and I prepared a paper on empirical age earnings profiles. By way of introduction, we wanted an approximate count of the number of articles in economics journals that had used the Mincer specification. Once we saw it was well over 1,000, we gave up and simply noted that fact.

Mincer's early work on human capital and earnings was interspersed with work on the labour force participation of married women (1960a, 1960b, 1962a). As for his work on wages and experience, it set the stage for a voluminous literature to follow. As noted, he introduced work on earnings profiles of women that included interrupted work careers (1974a, 1974b, 1978a, 1979, 1980). Although he subsequently added excellent work on wage growth and job mobility, the nemesis of economists – the minimum wage, and economic growth – I believe that the most outstanding is the 1978 *JPE* paper, ' Family Migration Decisions' , where he made the *ex post* obvious point regarding tried husband–wife movers, namely, in two-career families it is more difficult to find superior alternatives than it is for individuals or for one-career families. Moreover, the difficulty of finding superior alternatives increases as the specialization of the careers increases. As is true of most of his work, whether he pioneered or joined an existing literature, he greatly influenced what was to follow.

Mincer thought about and worked on important problems. He was original. You expected to learn any time you read a Mincer paper. Further, he always looked for applications: the theory had an empirical counterpart. Equally important, he was simply very good at what he did. He was an excellent colleague, teacher, and mentor to his doctoral students. He was also a great man. I am honoured to have known him.

Jacob Mincer retired from Columbia University in 1991. In 2002, the Institute for the Study of Labor (IZA) in Bonn awarded him the inaugural IZA Prize in Labor Economics. The prize was announced at his 80th birthday celebration hosted by Columbia University. In 2003 Mincer and Gary Becker were the inaugural recipients of the Society of Labor Economists (SOLE) Career Achievement Award. That award was then renamed the Jacob Mincer Award, in honour of the great man.

See Also

► Returns to Schooling
► Women's Work and Wages

Selected Works

1958. Investment in human capital and personal income distribution. *Journal of Political Economy* 66: 281–302.
1960a. Employment and consumption. *Review of Economics and Statistics* 42: 20–6.
1960b. Labor supply, family income and consumption. *Papers and Proceedings of the American Economic Association* 50: 574–83.
1962a. Labor force participation of married women. In *Aspects of Labor Economics,* ed. H.G. Lewis. Princeton: Princeton University Press.
1962b. On the job training: Costs, returns, and some implications. In *Investment in Human Beings,* 2, U-NB Conference, printed in *Journal of Political Economy* 70(Suppl): 50–79.
1974a. Family investment in human capital: Earnings of women, with S. Polacheck. *Journal of Political Economy* 82: 76–108.
1974b. *Schooling, Experience and Earnings.* New York: NBER.
1978a. Family migration decisions. *Journal of Political Economy* 86: 749–73.
1978b. (With S. Polachek.) Women's earnings reexamined. *Journal of Human Resources* 13: 118–34.
1979. (With H. Ofek.) Lifetime distribution of labor supply of married women. *Journal of Political Economy* 87: 197–202.
1980. Research in earnings and labor supply of women. In *Conference on Women in the Labor Market,* ed. C. Lloyd, E. Andrews and C. Gilroy. New York: Columbia University Press.

Bibliography

Becker, G.S. 1960. Underinvestment in college education? *American Economic Review* 50, 346–54. Reprint in *Problems of economic growth* ed. E. Phelps. New York: Norton, 1962.
Becker, G.S. 1962. Investment in human capital: A theoretical analysis. *Journal of Political Economy* 5: 9–49.
Schultz, T.W. 1960. Capital formation by education. *Journal of Political Economy* 6: 571–583.
Schultz, T.W. 1961a. Education and economic growth. In *Social forces influencing American education*, ed. N.B. Henry. Chicago: University of Chicago Press.
Schultz, T.W. 1961b. Investment in human capital. *American Economic Review* 51: 1–17.

Minimax

Jörg Stoye

Abstract

Minimax (Wald, *Statistical decision functions*. New York: Wiley, 1950) is the principle in statistical decision theory of minimizing worst-case risk. It is the subject of a rich literature in statistics and saw occasional normative application in economics. Minimax is related to the maximin expected utility model (Gilboa and Schmeidler, *J. Math. Econ.* 18:141–153, 1989) in economics, an model of ambiguity aversion that was recently used to analyse model uncertainty.

Keywords

Ambiguity; Decision theory; Econometrics; Estimation; Maxmin; Minimax; Minimax regret; Model uncertainty

JEL Classifications

D81

Minimax is the principle in statistical decision theory of optimizing worst-case outcomes. The minimax principle was first formalized by Wald in a sequence of papers culminating in Wald (1950). In statistics, minimax estimators or decision rules have since become the objects of a rich literature. Minimax is related to maxmin expected utility, a leading model of ambiguity aversion in economic theory that recently became prominent as a way to approach model uncertainty. While not the focus of this article, Rawls' (1999, first edition 1971) use of minimax as component of a normative theory of justice deserves mention.

Minimax in Statistics and Economics

This entry uses the same notation (essentially due to Wald 1950) as the one on econometrics and decision theory. A statistical experiment consists of a family of distributions $\{P_\theta : \theta \in \Theta\}$ over an outcome set Z. (Θ may be infinite dimensional, so the model underlying the experiment need not be parametric in the usual sense.) The decision maker must pick an act a from some feasible set A, possibly at random, after observing a draw z from P_θ. A complete contingent plan for this decision maker can be summarized by a decision function $\delta: Z \times [0, 1] \to A$, where $\delta(z, u)$ assigns treatment conditional on observation z and randomization u (u is normalized to be drawn from a uniform (0, 1) distribution). The decision maker incurs loss $L(a, \theta) \geq 0$; a decision rule's risk function $R(\delta, \theta) = \int\int L(a, \theta) dP_\theta du$ maps possible parameter values onto expected losses.

The question is which decision rule to pick, given that there are typically many undominated or admissible ones. Under minimax the answer is to minimize $\sup_{\theta \in \Theta} R(\delta, \theta)$: that is, worst-case risk. The best known alternative is Bayesianism, that is, to minimize $\int R(\delta, \theta) d\pi$, where π is a prior over Θ. A compromise between the two is Γ-minimax (Berger 1985), which imposes a set of priors Γ over Θ and then minimizes $\sup_{\pi \in \Gamma} \int R(\delta, \theta) d\pi$, the maximal expected risk over Γ. This nests standard minimax as the extremal case where Γ contains all possible priors over Θ.

It is instructive to compare this with the 'maxmin expected utility' model (Gilboa and Schmeidler 1989), in which a decision maker ranks acts according to $\min_{\pi \in \Gamma} \int u \circ f(s) d\pi(s)$. Here, f denotes an act and maps states of the world s into lotteries over ultimate outcomes x; Γ is a set of priors π; and u is a von Neumann–Morgenstern utility, thus $u \circ f(s) = \int U(x) df(s)$ for some utility function U. The notation is due to Anscombe and Aumann (1963) and is introduced in detail in this dictionary's entry on ambiguity.

These formalisms are related as follows. States of the world s correspond to parameter values θ. Losses $L(a, \theta)$ correspond to (negative) utility

evaluations of outcomes $U(x)$: that is, they are already expressed in utility terms. Because of this, outcomes themselves, as well as acts, do not have a direct analogue in Wald's setting. Conversely, risk functions correspond not to any of Anscombe and Aumann's primitives, but to so-called utility acts $u \circ f(s)$ that map states of the world into expected utilities and that play important roles in many axiomatic developments. Finally and most importantly, maxmin expected utility corresponds to Γ-minimax. The criterion function of classical minimax translates into the decision theoretic notation as $\min_{s \in S} u \circ f(s)$.

Foundations of Minimax

Foundations for minimax can be found in the axiomatic literature on decision theory. A natural starting point is Gilboa and Schmeidler's (1989) characterization of maxmin expected utility (and hence, Γ-minimax). The core insight behind this characterization concerns the following axioms for a preference ordering $\succcurlyeq$ over acts.

Independence
$f \succcurlyeq g$ iff $\alpha f + (1-\alpha)h \succcurlyeq \alpha g + (1-\alpha)h$ for all scalars $\alpha \in (0,1)$ and acts f, g, h; here, $\alpha f + (1-\alpha)h$ denotes a statewise probabilistic mixture of acts.

C-independence Like independence, but imposed only if h is constant: that is, h yields the same lottery in every state.

Uncertainty aversion $f \sim g$ implies $\alpha f + (1-\alpha)g \succcurlyeq f$ for all $\alpha \in (0,1)$.

The first of these axioms, von Neumann and Morgenstern (1947) independence, is crucial for characterizations of Bayesianism. Gilboa and Schmeidler (1989) replace it with the next two, weaker ones. Uncertainty aversion states that decision makers exhibit weak preference for mixtures, intuitively because these constitute a hedging of bets across states. Any such strict preference would violate independence. C-independence limits the potential for such violations by reinstating independence whenever the mixing act is constant, intuitively because mixing with constant acts cannot generate a hedge.

The resulting characterization leaves Γ unspecified. This is appropriate from a 'revealed preference' point of view because sets of beliefs are not directly observable, but users of the statistical minimax criterion might desire axiomatizations that imply the according – that is, the maximal – specification of Γ. These were originated by Milnor (1954) and modernized and made comparable to Gilboa and Schmeidler (1989) by Stoye (2006). Specifically, Γ can be made maximal by adding a symmetry axiom (Arrow and Hurwicz 1972; Cohen and Jaffray 1980) that excludes any prior weighting of states and thereby eliminates any vestige of Bayesianism.

Finding Minimax Rules

No universal method for finding minimax decision rules exists, but a number of helpful ones are detailed in any statistics textbook. See for example Berger (1985) for an overview and Ferguson (1967) for an advanced treatment. A technique of special interest to economists is direct application of game theory (Wald 1945).

Let π* be a prior and δ* a decision rule such that (i) $\int R(\delta^*, \theta) d\pi^* \geq \int R(\delta^*, \theta) d\pi$ for any prior π over Θ, thus π^* maximizes risk given δ^*; and (ii) δ^* is the Bayes rule relative to π^*. Then δ^* achieves minimax risk. π^* is also called a least favorable prior, and it can be instructive to think of (δ^*, π^*) as Nash equilibrium of a fictitious zero-sum game between the decision maker and a malicious Nature. The minimax theorem (von Neumann 1928) gives conditions under which π* exists; subsequent existence results for Nash equilibria imply other such conditions. The technique can also be extended to cases where π^* fails to exist; specifically, δ^* is minimax if there exists a sequence (δ_n, π_n) such that δ_n is Bayes relative to π_n and $\sup_{\theta \in \Theta} R(\delta^*, \theta) \leq \lim_{n \to \infty} \int R(\delta_n, \theta) d\pi_n < \infty$. Some other

techniques for finding minimax rules – for example, minimaxity of a constant-risk Bayes rule – are corollaries.

As an example, let Z be binomial with parameters (θ, n) and let $L(a, \theta) = (a - \theta)^2$. Then $\delta(z) = (z + \sqrt{n}/2)/(n + \sqrt{n})$ can be shown to have constant risk and to be Bayes if π^* is a Beta $(\sqrt{n}/2, \sqrt{n}/2)$ distribution, hence it is a minimax estimator. Note that the sample analogue of θ, z/n, might appear a more natural estimator; δ shrinks it toward 1/2.

In more involved problems, finding exact minimax rules may not be feasible, and one may have to resort to asymptotic analysis (Le Cam 1986). A classic result is that under certain conditions, Bayes as well as maximum likelihood estimators are locally asymptotically minimax.

Applications

A famous early application of minimax to estimation is Hodges and Lehmann (1950). A sizeable literature developed from this and is surveyed in the textbooks mentioned above. Chamberlain (2008) applies minimax analysis to an instrumental variables model; a core result is that under normality and other conditions, maximum likelihood is (finite sample) minimax for some parameters. Chamberlain (2000) applies minimax to portfolio choice problems. Robust control in macroeconomics has a maxmin expected utility interpretation; see Hansen et al. (2006) and this dictionary's entry on 'model uncertainty'. In economic theory, maxmin expected utility is an early benchmark in the large literature on ambiguity aversion; see this dictionary's entry on 'ambiguity and ambiguity aversion'.

Criticisms of Minimax

Criticisms of minimax centre on the facts that it may be perceived as extremely conservative and that it may optimize against an implausible prior. For example, π^* in the above example is much concentrated near 1/2. (Intuitively, values of θ close to 1/2 are unfavourable because they imply a large variance of the signal.) The sample analogue of θ accordingly underperforms against the minimax estimator if θ is indeed close to 1/2, but outperforms it by a much greater margin for θ near 0 or 1 and is generally considered more attractive. Also, minimax estimators need not be admissible; while admissible minimax rules exist under regularity conditions, the techniques described above might not identify them. Furthermore, it is easy to construct decision problems in which minimax decision rules ignore available data (Savage 1954), and in economics, recent work on treatment choice uncovered natural examples of this (Manski 2004). Users who are comfortable with priors can avoid some of these criticisms by using the Bayesian or Γ-minimax criteria, taking care to specify reasonable priors. Economists looking for non-Bayesian approaches recently explored minimax regret as an alternative (Manski 2004 and other references in the 'minimax regret' entry). Finally, it is not obvious how to adapt the minimax principle to dynamic decision problems; see the entry on ambiguity aversion.

M

See Also

- ▶ Ambiguity and Ambiguity Aversion
- ▶ Decision Theory in Econometrics
- ▶ Minimax Regret
- ▶ Model Uncertainty

Bibliography

Anscombe, F.J., and R.J. Aumann. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34: 199–205.

Arrow, K.J., and L. Hurwicz. 1972. An optimality criterion for decision-making under ignorance. In *Uncertainty and expectations in economics: Essays in honour of G.L.S. Shackle*, ed. C.F. Carter and J.L. Ford. London: Blackwell.

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*, 2nd ed. New York: Springer.

Chamberlain, G. 2000. Econometric applications of maxmin expected utility. *Journal of Applied Econometrics* 15: 625–644.

Chamberlain, G. 2008. Decision theory applied to an instrumental variables model. *Econometrica* 75: 609–652.

Cohen, M., and J.-Y. Jaffray. 1980. Rational behavior under complete ignorance. *Econometrica* 48: 1281–1299.
Ferguson, T. 1967. *Mathematical statistics: A decision theoretic approach*. New York: Academic Press.
Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18: 141–153.
Hansen, L.P., T.J. Sargent, G. Turmuhambetova, and N. Williams. 2006. Robust control and model misspecification. *Journal of Economic Theory* 128: 45–90.
Hodges, J.L., and E.L. Lehmann. 1950. Some problems in minimax point estimation. *Annals of Mathematical Statistics* 21: 182–197.
Le Cam, L. 1986. *Asymptotic methods in statistical decision theory*. New York: Springer.
Manski, C.F. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72: 1221–1246.
Milnor, J. 1954. Games against nature. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis. New York: Wiley.
Rawls, J. 1999. *A theory of justice*, Rev. ed. Cambridge, MA: Harvard University Press.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Stoye, J. 2006. *Statistical decisions under ambiguity*, Discussion paper. New York: New York University.
von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.
von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press.
Wald, A. 1945. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics* 46: 265–280.
Wald, A. 1950. *Statistical decision functions*. New York: Wiley.

Minimax Regret

Jörg Stoye

Abstract

Minimax regret (Savage, Journal of the American Statistical Association 46, 55–67, 1951) is the principle of optimizing worst-case loss relative to some measure of unavoidable risk. In statistical decision theory, it provides a non-Bayesian alternative to minimax. It differs from minimax by fulfilling von Neumann–Morgenstern independence but exhibiting menu dependence. Minimax regret has seen occasional use in statistics, and implausible implications of minimax in certain economic problems recently led to its reconsideration by economists.

Keywords

Decision theory; Econometrics; Estimation; Maxmin; Minimax; Minimax regret; Model uncertainty

JEL Classifications

D81

Minimax regret is the principle in statistical decision theory of optimizing worst-case efficiency loss relative to an ex post optimal decision. It was originally proposed in Savage's (1951) review of Wald (1950). In fact, Savage misinterpreted Wald (1950) and took it that he had proposed minimax regret rather than minimax; this was clarified in Savage (1954). The principle saw occasional use in statistics and machine learning (Das Gupta and Studden 1991; Droge 1998; Foster and Vohra 1999) and recently enjoyed some revival in economics and econometrics, especially with regard to treatment choice (see below for references).

Definition and Foundations

This article uses the same notation as the entries on minimax and on econometrics and decision theory; see either for elaborations. A minimax regret statistical decision rule minimizes (over D, the set of feasible decision rules)

$$\sup_{\theta \in \Theta} \left\{ R(\delta, \theta) - \inf_{\delta * \in D} R(\delta^*, \theta) \right\},$$

where R denotes a risk function. Thus, $R(\delta, \theta)$ is the expected loss incurred by decision rule δ as function of some unknown parameter value θ, and $\inf_{\delta^* \in D} R(\delta^*, \theta)$ indicates the lowest expected loss achievable given θ. Minimax regret differs from standard minimax by considering not loss in and

of itself, but excess loss relative to this unavoidable risk. Intuitively, it thereby optimizes not against parameter values that are unfavourable to any decision rule, but against ones where a decision rule can cause great damage. Unlike with minimax, D enters the criterion function, hence minimax regret is menu dependent. Expanding D can affect the preferred decision rule even if the newly available rules are themselves unattractive.

Important variations are as follows. First, the preceding criterion is prior-free: the supremum is taken over all possible parameter values, without any concern for prior probabilities. The Γ-minimax regret criterion (Berger 1985) takes the supremum with respect to a set of priors over Θ and thereby allows for compromises with Bayesianism. Second, the benchmark $\inf_{\delta^* \in D} R(\delta^*, \theta)$ could be defined via the ex post best among a subset D' of decision rules, a well-known example being Hannan regret (Cesa-Bianchi and Lugosi 2006; Hannan 1957). A close relative of minimax regret which enjoys some popularity in computer science is the competitive ratio, defined by taking the ratio rather than the difference to unavoidable risk (Borodin and El-Yaniv 1998).

The prior-less minimax regret preference ordering was axiomatized by Milnor (1954) and Stoye (2006). However, menu dependence implies that preferences over decision rules that are not in fact chosen lack a behavioural or 'revealed preference' interpretation. Hayashi (2008) provides a revealed preference characterization of the Γ-minimax regret choice correspondence. Stoye (2007b) subsequently unifies the literature and considers prior-less minimax, Γ-minimax and Hannan regret. In either framework, the core message is that the trade-off between minimax and minimax regret can be cast as choice among two well-known axioms. Minimax avoids the aforementioned menu dependence but violates von Neumann–Morgenstern independence (for which see the entry on minimax); minimax regret fulfils independence but is menu dependent.

Mathematically, minimax regret is minimax with a respecified risk function, so remarks on finding minimax regret rules mirror the relevant remarks for minimax. Recent applications of game theory to identify finite sample minimax regret rules include Schlag (2007) and Stoye (2007c, d). Asymptotic minimax regret efficiency is treated by Hirano and Porter (2006).

Applications

Minimax regret coincides with minimax loss if $\inf_{\delta^* \in D} R(\delta^*, \theta)$ is constant on Θ, as in the estimation example given in the entry on minimax. The criteria differ in the following, simple application. A decision maker must assign one of two treatments $t \in \{0, 1\}$ to some population; treatments induce random outcomes (Y_0, Y_1) supported on $\{success, failure\}$. Treatment choice can condition on observations from a simple random sample of N subjects, half of whom were assigned to treatment $t = 0$. Then the no-data rule that always assigns treatment 0 is minimax, as is any other decision rule. This is because any decision rule's risk is maximized – and all these maxima are identical – if both treatments induce only failures. The more natural rule that assigns everybody to the treatment that scores more successes in the sample (with even tie-breaking) is asymptotically (Hirano and Porter 2008) as well as finite sample (Canner 1970) minimax regret efficient, and essentially uniquely so (Stoye 2007d).

The example illustrates a classic criticism (Savage 1951) of minimax, namely that its 'ultra-pessimism' can lead to complete ignorance of data, as well as how minimax regret may avoid the problem. Extensions of this application are analyzed in Brock (2006), Manski (2004, 2007), Stoye (2007a, c, d), and Schlag (2007). Empirical applications of minimax regret to treatment choice are found in Eozenou et al. (2006), Manski (2008) and Stoye (2009). In other applications to economics, Schlag (2003) brings minimax regret to bandit problems; Hansen (2005) evaluates kernel density estimators in terms of minimax regret; Bergemann and Schlag use minimax regret (2008) and Γ-minimax regret (2007) to analyse monopoly pricing; Chamberlain (2000) applies Γ-minimax regret to portfolio choice problems; and Hart and Mas-Colell (2001) use Hannan regret to evaluate learning rules.

Criticisms

The menu dependence of minimax regret has attracted criticism at least since Chernoff (1954). Other criticisms mirror those of the minimax principle, namely that minimax regret may implicitly optimize against unreasonable priors. It is worth noting that while minimax regret avoids no-data rules in natural examples, examples that go the other way can be constructed (Parmigiani 1992). A natural example in which both principles inform no-data rules occurs if one modifies the above application by conditioning on a continuous covariate (Stoye 2007d).

See Also

- ▶ Decision Theory in Econometrics
- ▶ Minimax

Bibliography

Bergemann, D., and K.H. Schlag. 2007. *Robust monopoly pricing*, Cowles Foundation Discussion Paper 1527R. New Haven: Yale University.

Bergemann, D., and K.H. Schlag. 2008. Pricing without priors. *Journal of the European Economic Association* (*Papers and Proceedings*) 6: 560–569.

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*, 2nd ed. New York: Springer.

Borodin, A., and R. El-Yaniv. 1998. *Online computation and competitive analysis*. Cambridge/New York: Cambridge University Press.

Brock, W.A. 2006. Profiling problems with partially identified structure. *Economic Journal* 92: F427–F440.

Canner, P.L. 1970. Selecting one of two treatments when the responses are dichotomous. *Journal of the American Statistical Association* 65: 293–306.

Cesa-Bianchi, N., and G. Lugosi. 2006. *Prediction, learning, and games*. Cambridge: Cambridge University Press.

Chamberlain, G. 2000. Econometrics and decision theory. *Journal of Econometrics* 95: 255–283.

Chernoff, H. 1954. Rational selection of decision functions. *Econometrica* 22: 422–443.

Das Gupta, A., and W. Studden. 1991. Robust Bayesian experimental designs in normal linear models. *Annals of Statistics* 19: 1244–1256.

Droge, B. 1998. Minimax regret analysis of orthogonal series regression estimation: Selection versus shrinkage. *Biometrika* 85: 631–643.

Eozenou, P., J. Rivas, and K.H. Schlag. 2006. Minimax regret in practice: Four examples on treatment choice. Discussion paper. Florence: European University Institute.

Foster, D., and R. Vohra. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29: 7–36.

Hannan, J. 1957. Approximation of Bayes risk in repeated play. In *Contributions to the theory of games*, vol. III, ed. M. Dresher, A.W. Tucker, and P. Wolfe. Princeton: Princeton University Press.

Hansen, B.E. 2005. Exact mean integrated squared error of higher-order kernels. *Econometric Theory* 21: 1031–1057.

Hart, S., and A. Mas-Colell. 2001. A general class of adaptive strategies. *Journal of Economic Theory* 98: 26–54.

Hayashi, T. 2008. Regret aversion and opportunity-dependence. *Journal of Economic Theory* 139: 242–268.

Hirano, K., and J. Porter. 2008. Asymptotics for statistical treatment rules. Discussion paper. Tucson: University of Arizona.

Manski, C.F. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72: 1221–1246.

Manski, C.F. 2007. Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics* 139: 105–115.

Manski, C.F. 2008. *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.

Milnor, J. 1954. Games against nature. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis. New York: Wiley.

Parmigiani, G. 1992. Minimax, information and ultra-pessimism. *Theory and Decision* 33: 241–252.

Savage, L.J. 1951. The theory of statistical decision. *Journal of the American Statistical Association* 46: 55–67.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Schlag, K.H. 2003. How to minimize maximum regret in repeated decisions. Discussion paper. Florence: European University Institute.

Schlag, K.H. 2007. Eleven: Designing randomized experiments under minimax regret. Discussion paper. Florence: European University Institute.

Stoye, J. 2006. Statistical decisions under ambiguity. Discussion paper. New York: New York University.

Stoye, J. 2007a. Minimax regret treatment choice with incomplete data and many treatments. *Econometric Theory* 23: 190–199.

Stoye, J. 2007b. Axioms for minimax regret choice correspondences. Discussion paper. New York: New York University.

Stoye, J. 2007c. Minimax regret treatment choice with finite samples and missing outcome data. In *Proceedings of the fifth international symposium on imprecise probability: Theories and applications*, ed. G. de Cooman, J. Veinarová, and M. Zaffalon. Prague.

Stoye, J. 2007d. Minimax regret treatment choice with finite samples. Discussion paper. New York: New York University.
Stoye, J. 2009. Minimax regret treatment choice with missing data: An application to young offenders. *Journal of Statistical Theory and Practice*, forthcoming.
Wald, A. 1950. *Statistical decision functions*. New York: Wiley.

Minimum Wages

Donald O. Parsons

Abstract

The minimum wage, the lowest wage rate legally payable by employers to workers, derives support from concern about the equity of market processes. Because employment may fall in response to an increase in the minimum wage and because the majority of low-wage workers do not come from families in poverty, the minimum wage may have modest benefits as a poverty reduction tool. While there are variations across studies, evidence from the United States suggests that the economy-wide employment effects of wage minimums at the levels at which they have been implemented in the United States are negative but not large.

Keywords

Envelope theorem; Hours worked; Labour market participation; Labour supply; Minimum wages; Poverty alleviation; Unemployment

JEL Classifications

J3

The term minimum wages refers to various legal restrictions on the lowest wage rate payable by employers to workers. Until relatively recently, wage floors usually had a very specific focus; in Great Britain and the United States, for example, minimum wages were initially limited to women and children. Only following the Great Depression were such laws extended systematically to the general work force in many industrial and industrializing economies. The minimum wage restrictions were often industry specific, in France for example, extensions of trade union legislation (Rosa 1981). In the United States, industry-specific wage restrictions were held to be unconstitutional; in 1938 a uniform national minimum wage rate was established for non-farm, non-supervisory personnel under the Fair Labor Standards Act.

Subsequently, coverage was extended to the bulk of the labour force.

The social appeal of minimum wage legislation appears to be strong, its intuitive base rooted in concern about the equity of market processes. Dissatisfaction with the share of production allocated to the least able members of the work force is prevalent even among individuals impressed with the enormous capacity of the market system to organize productive activity. An obvious solution to this problem, and one that can be implemented with a modest government budget commitment for statute enforcement, is to redefine the wage structure politically to achieve a socially preferable distribution of income. Although the political interests that have formed the most prominent support for minimum wage legislation may have had less socially oriented goals, for example, Colberg (1960) and Silberman and Durden (1978), broad public support for such legislation is, I believe, based on this equity issue and it is usually against the social criterion of poverty reduction that minimum wage legislation has been judged.

Stigler (1946) provides the classic discussion of the potential deficiencies of minimum wage legislation as an antipoverty device; employment may fall more than in proportion to the wage increase from the minimum, thereby reducing earnings: wage rates in uncovered sectors may decrease more than those in the covered sector rise as the uncovered sector is forced to absorb the workers released by the covered sector: the impact of the legislation on family income distribution may be perverse unless the fewer but better

jobs are allocated to members of needy families rather than to low-wage workers, most obviously teenagers, from wealthier families. A crucial insight by economists is that minimum wage legislation alters the opportunity set of the least able but does not unambiguously expand it. The legal restriction that employers cannot pay less than a specified wage is equivalent to the legal stipulation that workers cannot work at all in the protected sector unless they find employers willing to hire them at that wage. Much of the progress in the analysis of minimum wage effects in the last several decades has focused on the theoretical and empirical modelling needed to assess the welfare implications of this altered opportunity set.

As the theoretical modelling of the low-wage labour market has become more complete, theoretical predictions of minimum wage law effects have, unfortunately, become qualitatively ambiguous. Most models have been designed to capture the major features of minimum wage legislation in the United States, a uniform wage minimum covering a portion of a competitive economy. The principal implication of such models is that employment in the covered sector will fall with the establishment of an effective minimum wage. If labour supply is inelastic, these disemployed workers will seek and presumably find employment in the uncovered sector. The wages and well-being of workers in the uncovered sector might be expected to fall as that sector is forced to absorb additional workers (Stigler 1946; Welch 1974, 1978). Johnson (1969) demonstrates, however, that in a general equilibrium framework with two factors (labour and capital) the well-being of uncovered workers could in fact rise. If the covered sector is sufficiently capital intense and faces a sufficiently high demand elasticity, the quantity of capital released as the covered sector contracts could potentially increase the well-being of workers in the uncovered sector. The introduction of an elastic labour supply function (and implicitly or explicitly some valuable non-market activity) suggests additional parameters that must be estimated before theoretical considerations can be brought to bear on the assessment of the minimum wage (Welch 1974).

The modelling of minimum wage effects on unemployment and labour force participation is more complex than on employment, requiring careful specification of the search process (Mincer 1976). The effect of a minimum wage on unemployment, for instance, depends critically on the queuing method required to secure high paying jobs and on the optimal search strategy induced by this hiring regime. If, for example, workers must wait in a union hall to secure jobs in the covered sector, the extent of unemployment will be quite different than if they can maintain their places in the queue for covered employment while in an uncovered sector job or while out of the labour force entirely.

Before turning to the empirical evidence on minimum wage effects, a brief comment on compliance is warranted. Legal compliance with the minimum wage laws in the United States appears to be surprisingly high (Ashenfelter and Smith 1979). Effective evasion of small minimum wage restrictions, however, is probably quite high since wages are only a portion of the employment compensation package (Wessels 1980). Non-wage benefits such as paid vacations are almost completely fungible. Indeed, the envelope theorem would suggest that modest adjustments among components in the total compensation package could be made without affecting employer costs or, equally important, worker welfare. Larger minimum wage restrictions would presumably raise covered worker welfare and employer costs, but not at the rate suggested by the wage-only compensation models.

Among other adjustments employers could make to an increase in the wage minimum would be an increase in effort demands or a reduction in the convenience (or number) of scheduled work hours. Perhaps of greater concern to economists is the potential for a reduction in the provision of on-the-job training to the young. The adverse training effects of legal minimum wages appear to be significant (Leighton and Mincer 1981; Hashimoto 1982), although perhaps partly offset by increased schooling in a broader picture (Mattila 1981).

Clearly the effect of minimum wage laws on the wages and well-being of the labour force must

be resolved empirically, either by estimation of the parameters in the theoretical models or by direct estimation of labour market effects. The latter approach has been the most common. Unfortunately, the evidence for the United States labour market (for which such estimation is most prevalent) is not as useful as one might hope. The political equilibrium in the United States has apparently kept the legal wage minimum relatively low. Only in a few circumstances has the minimum been so large as to induce major industrial contractions, for example in the South in the early years of the legislation (Colberg 1960), and in Puerto Rico, most dramatically in that same period (Reynolds and Gregory 1965). For most of the more recent period, the wage minimum has been primarily limited in impact to teenagers of both sexes and to adult females (Kneisner 1981), both of which groups have significant non-market alternatives subject to their own exogenous forces.

The empirical literature on employment effects of the legal minimum wage in the United States suggests that the economy wide employment effects of wage minimums at recent levels are negative but not large (Eccles and Freeman 1982; Brown et al. 1982). Most estimates are bounded by employment elasticities of minus 1 (a reduction in employment equiproportional to the increase in the wage minimum) and zero. Brown, Gilroy and Kohen argue for an estimate towards the zero portion of that range. The effects may, however, not be constant over a wider range of minimum wage restrictions; as the potential for substitution within the total compensation package is reduced, the employment effects will almost surely increase. Certainly minimum wage restrictions that are 'large' relative to customary wages appear to have very large effects, whether considered regionally (again Colberg 1960; Reynolds and Gregory 1965), or by economic sector (Fleisher 1981).

Highly visible work by Card and Krueger (1994, 1995) has focused on 'natural experiments' generated by changes in the minimum wage. In 1992 the minimum wage was increased in New Jersey. Card and Krueger estimated the effect of the minimum wage on employment in fast-food restaurants in New Jersey compared with neighbouring Pennsylvania, where there was no increase in the minimum wage, and found that employment increased in New Jersey relative to Pennsylvania.

Kennan (1995) discussed this work and potential explanations and subsequent research by Neumark and Wascher (2000) questioned their results.

For most individuals not directly involved in buying or selling low skilled labour, the critical empirical question is not the magnitude of the employment effects of minimum wages but rather the effect on income poverty. Obviously large negative employment effects would suggest that the antipoverty effects of the minimum wage are small or possibly even perverse. Direct empirical studies of antipoverty effects (Gramlich 1976; Parsons 1980) indicate, however, that the antipoverty effects in the United States would be quite modest even if employment effects were zero. The great majority of low-wage workers do not come from families in poverty. Moreover, the groups primarily affected, teenagers and low-skilled adult females, are predominantly part-time workers and any wage-rate effect on earnings and income is strictly proportional to hours worked. Even a fully effective wage minimum with no offsetting employment adjustment would provide little relief to poverty-level families (Parsons 1980). Negative employment effects simply enhance other fundamental limitations of minimum wage legislation as a poverty programme.

Wage rate restrictions alone appear to be an unsatisfactory solution to social concerns about labour market outcomes. Politically manipulating the price system seems like a direct and inexpensive method of assisting the disadvantaged. Almost surely it is not. Employment opportunities and the factors that limit labour market participation must be considered as well as wage rates if market outcomes are to be supplanted in a socially satisfactory way for low-skilled workers.

See Also

► Labour Economics

Bibliography

Ashenfelter, O., and R.S. Smith. 1979. Compliance with the minimum wage law. *Journal of Political Economy* 87: 333–350.

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

Card, D., and A.B.. Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84: 772–793.

Card, D., and A.B.. Krueger. 1995. *Myth and measurement: The new economics of the minimum wage*. Princeton: Princeton University Press.

Colberg, M.R. 1960. Minimum wage effects on Florida's economic development. *Journal of Law and Economics* 3: 106–117.

Eccles, M., and R.B. Freeman. 1982. What! Another minimum wage study? *American Economic Review* 94: 226–232.

Fleisher, B.M. 1981. *Minimum wage regulation in retail trade*. Washington, DC: American Enterprise Institute.

Gramlich, E.M. 1976. Impact of minimum wages on other wages, employment, and family incomes. *Brookings Papers on Economic Activity* 1976 (2): 409–451.

Hashimoto, M. 1982. Minimum wage effects on training on the job. *American Economic Review* 72: 1070–1087.

Johnson, H.J. 1969. Minimum wage laws: A general equilibrium analysis. *Canadian Journal of Economics* 2: 599–604.

Kennan, J. 1995. The elusive effects of minimum wages. *Journal of Economic Literature* 33: 1950–1965.

Kneisner, T.J. 1981. The low-wage workers: Who are they? In Rottenberg (1981).

Leighton, L., and J. Mincer. 1981. The effects of minimum wages on human capital formation. In Rottenberg (1981).

Mattila, J. 1981. The impact of minimum wages on teenage schooling and part-time full-time employment of youths. In Rottenberg (1981).

Mincer, J. 1976. Unemployment effects of minimum wages. *Journal of Political Economy* 84: S87–S104.

Neumark, D., and W. Wascher. 2000. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. *American Economic Review* 90: 1362–1396.

Parsons, D.O. 1980. *Poverty and the minimum wage*. Washington, DC: American Enterprise Institute.

Reynolds, L.G., and P. Gregory. 1965. *Wages, productivity, and industrialization in Puerto Rico*. Homewood: Richard D. Irwin.

Rosa, J.J. 1981. The effects of minimum wage regulation in France. In Rottenberg (1981).

Rottenberg, S., ed. 1981. *The economics of legal minimum wages*. Washington, DC: American Enterprise Institute.

Silberman, J., and G.C. Durden. 1978. Determining legislative preferences on the minimum wage: An economic approach. *Journal of Political Economy* 84: 317–329.

Stigler, G.J. 1946. The economics of minimum wage legislation. *American Economic Review* 36: 358–365.

Welch, F. 1974. Minimum wage legislation in the United States. *Economic Inquiry* 12: 285–318.

Welch, F. 1978. *Minimum wages: Issues and evidence*. Washington, DC: American Enterprise Institute.

Wessels, W.J. 1980. The effect of minimum wages on fringe benefits: An expanded model. *Economic Inquiry* 18: 293–213.

Minsky Crisis

L. Randall Wray

Abstract

This entry examines the approach of Hyman P. Minsky to financial crisis. Minsky famously developed an 'investment theory of the cycle and a financial theory of investment'. His thesis was that, over the course of the cycle, behaviour changes in such a way that financial fragility develops. This makes a financial crisis more likely. When the global financial crisis hit in 2008, many commentators returned to the theories of Minsky, calling it a 'Minsky crisis' or a 'Minsky moment'. This entry agrees that Minsky deserves credit for identifying the processes that led up to the crisis. However, it is not sufficient to narrowly constrain the analysis to the transition that occurred over the past decade or so. Beginning in the 1980s and through to his death in 1996, Minsky had been arguing that a new form of capitalism had appeared, which he called 'money manager capitalism'. In important respects it reproduced the conditions that Hilferding had called 'finance capitalism' in the early 20th century – a form of capitalism that collapsed into the Great Depression. What Minsky was arguing was that an extremely unstable form of capitalism had emerged – one based on what is often called financialisation of the economy. He (rightly) feared that it would ultimately

lead to a great crash. The rest of the entry looks at Minsky's proposals for reforms that would help to promote stability. Yet, as Minsky always said, stability is destabilising.

Keywords

Financial instability hypothesis; Global financial crisis; Hyman Minsky; Money manager capitalism; Self-Regulating markets; Stability is destabilizing

JEL Classifications

B22; B25; B26; B52; E02; E11; E12; E44; G01; O11

Introduction

Stability is destabilizing. Those three words capture in a concise manner the insight that underlies Minsky's analysis of the transformation of the economy over the entire post-war period. The basic thesis is that the dynamic forces of the capitalist economy are explosive so that they must be contained by institutional ceilings and floors – part of the 'safety net'. However, to the extent that the constraints successfully achieve some semblance of stability, that will change behaviour in such a manner that the ceiling will be breached in an unsustainable speculative euphoria. If the inevitable crash is cushioned by the institutional floors, the risky behaviour that caused the boom will be rewarded. Another boom will build, and its crash will again test the safety net. Over time, the crises become increasingly frequent and severe until finally 'it' (a great depression with a debt deflation) becomes possible.

While Minsky's 'financial instability hypothesis' is fundamentally pessimistic, it is not meant to be fatalistic (Minsky 1975, 1982, 1986) According to Minsky, policy must adapt as the economy is transformed. The problem with the stabilizing institutions that had been put in place in the early post-war period is that they no longer served the economy well by the 1980s, as they had not kept up with the evolution of financial institutions and practices. Further, they had been purposely degraded and even in some cases dismantled, often on the erroneous belief that 'free' markets are self-regulating. Indeed, that became the clarion call of most of the economics profession after the early 1970s, based on the rise of 'new' classical economics with its rational agents and instantaneously clearing markets and the 'efficient markets hypothesis' that proclaimed prices fully reflect all information about 'fundamentals'. Hence, not only had firms learned how to circumvent regulations and other constraints, but policymakers had removed regulations and substituted 'self-regulation' in place of government oversight.

From his earliest writings in the late 1950s to his final papers written before his death in 1996, Minsky always analyzed the financial innovations of profit-seeking firms that were designed to subvert New Deal constraints. For example, he was one of the first economists to recognize how the development of the federal funds market had already reduced the Fed's ability to use reserves to constrain bank lending, while at the same time 'stretching' liquidity because banks would have fewer safe and liquid assets should they need to unwind balance sheets (Minsky 1957). And much later, in a remarkably prescient piece in 1987, Minsky had foreseen the development of securitization (to move interest rate risk off bank balance sheets while reducing capital requirements) that would later be behind the global financial crash of 2007 (published as Minsky 2008) At the same time, Minsky continually formulated and advocated policy to deal with these new developments. Unfortunately, his warnings were largely ignored by the profession and by policymakers – until it was too late.

Minsky's Theory of the Business Cycle

In the introduction I focused on long-term transformations because too often Minsky's analysis is interpreted as a theory of the business cycle. There have even been some analyses that attempted to 'prove' Minsky wrong by applying his theory to

data from one business cycle. Further, the global crisis that began in 2007 has been called the 'Minsky moment' or a 'Minsky crisis'. As I will discuss, I agree that this crisis does fit with Minsky's theory, but I object to analyses that begin with, say, 2004 – attributing the causes of the crisis to changes that occurred over a handful of years that preceded the collapse. Rather, I argue that we should find the causes of the crisis in the transformation that began in 1951. We will not understand the crisis if we begin with a US real estate boom fueled by lending to subprime borrowers. That will be the topic of the next section.

Now, Minsky *did* have a theory of the business cycle (see Papadimitriou and Wray (1998) for a summary of Minsky's approach). He called it 'an investment theory of the cycle and a financial theory of investment'. He borrowed the first part of that from Keynes: investment is unstable and tends to be the driver of the cycle (through its multiplier impact). Minsky's contribution was the financial theory of investment, with his book *John Maynard Keynes* (1975) providing the detailed exposition. In brief, investment is financed with a combination of internal and external (borrowed) funds. Over an expansion, success generates a greater willingness to borrow, which commits a rising portion of expected gross profits (Minsky called it gross capital income) to servicing debt. This exposes the firm to greater risk because if income flows turn out to be less than expected, or if finance costs rise, firms might not be able to meet those debt payment commitments. There is nothing inevitable about that, however, because Minsky incorporated the profits equation of Michal Kalecki in his analysis: at the aggregate level total profits equal investment plus the government's deficit plus net exports plus consumption out of profits and less saving out of wages (Minsky 1986). The important point is that all else being equal, higher investment generates higher profits at the aggregate level. This can actually make the system even more unstable, because if profits continually exceed expectations, making it easy to service debt, then firms will borrow even more.

This then leads to Minsky's famous categorization of financial positions: a hedge unit can meet payment commitments out of income flow; a speculative unit can only pay interest but must roll over principal; and a Ponzi unit cannot even make the interest payments so must 'capitalize' them (borrowing to pay interest). (In his classification of 'Ponzi finance', Minsky borrowed the name of a famous fraudster, Charles Ponzi, who ran a 'pyramid' scheme – in more recent times, Bernie Madoff ran another pyramid that failed spectacularly). Over a 'run of good times', firms (and households) are encouraged to move from hedge to speculative finance, and the economy as a whole transitions from one in which hedge finance dominates to one with a greater weight of speculative finance. Eventually some important units find they cannot pay interest, driving them to Ponzi finance. Honest bankers do not like to lend to Ponzi units because their outstanding debt grows continually unless income flows eventually rise. When the bank stops lending, the Ponzi unit collapses. Following Irving Fisher, Minsky then described a 'debt deflation' process: collapse by one borrower can bring down his creditors, who default on their own debts, generating a snowball of defaults. Uncertainty and pessimism rise, investment collapses and through the multiplier income and consumption also fall, and we are on our way to a recession.

But Minsky did not mean to imply that all financial crises lead to recessions, nor that all recessions result from the transition to speculative and Ponzi finance. The Federal government in the post-war period was big – 20–25% of the economy versus only 3% on the verge of the Great Depression. This meant that government itself could be both stabilizing and destabilizing. Countercyclical movement of its budget from surplus in a boom to deficit in a slump would stabilize income and profits (recall from the Kalecki accounting identity above that government deficits add to profits). A rising deficit could potentially offset the effects of falling investment, and, indeed, over the post-war period that helped to cushion every recession. However, it is also possible for the government to cause a downturn by cutting spending – as it did in the demobilization from the Second World War. And if the budget is excessively biased toward surplus when the economy grows, it will generate 'fiscal drag' that

removes household income and profits of firms – causing a recession. For that reason, a recession could occur well before the private sector is dominated by speculative and Ponzi positions. (Note that an economy that moves toward current account deficits when it grows robustly – such as the USA – will suffer an additional 'headwind' that sucks income and profits from domestic households and firms.)

In addition to the 'big government', the post-war period also had what Minsky called the 'big bank' – the Federal Reserve. The Fed plays a number of roles: it sets interest rates, it regulates and supervises banks, and it acts as lender of last resort. Generally, it moves interest rates in a pro-cyclical manner (raising them in expansion and lowering them in recession), which is believed by many orthodox economists to be stabilizing. Like many heterodox economists, Minsky doubted that spending is very interest-sensitive: in a boom, raising rates by a moderate amount will not curb enthusiasm, and in a bust, even very low interest rates cannot overcome pessimism. In addition, Minsky emphasized the impact of interest rates on financial fragility: raising rates in a boom would increase finance costs and hasten the transition to speculative and Ponzi financial positions, hence, to the extent that tight monetary policy 'works', it does so by inducing a financial crisis. Thus, Minsky rejected the notion that the Fed can use interest rates to 'fine tune' the economy.

But lender of last resort policy was viewed by Minsky as essential – it would stop a bank run and would help to put a floor to asset prices, attenuating the debt deflation process discussed above. If the Fed lends to a troubled financial institution, it does not have to sell assets to try to cover demands by creditors for redemption. For example, if depositors are demanding cash withdrawal, in the absence of a lender of last resort the bank would have to sell assets to raise the cash required; this is normally difficult for assets such as loans, and nearly impossible to do in a crisis. So the Fed lends the reserves to cover withdrawals.

In sum, the intervention of the big bank and the big government helps to prevent a financial crisis from turning into a deep downturn. The big government's deficit puts a floor to falling income and profits, and the big bank's lending relieves pressure in financial markets (Minsky 1986). A financial crisis can even occur without setting off a recession – a good example was the 1987 stock market crash, in which the Fed quickly intervened with the promise that it would lend reserves to market participants to stop necessitous selling of stocks to cover positions. No recession followed the crash – unlike the October 1929 crash, in which margin calls forced sales of stocks. And the big government deficits kept profits flowing in 1987, again unlike 1929 when the government's budget was far too small to make up for collapsing investment.

Unfortunately, most Fed policy over the post-war period involved reducing regulation and supervision, promoting the natural transition to financial fragility. From Minsky's perspective, this was a dangerous combination. While the big bank and the big government reduced the fall-out of crisis, the move to 'self-regulation' by financial institutions and markets made riskier behaviour possible. As the fear of failure was attenuated by a government safety net, perceived risk was lowered. Chairman Ben Bernanke (2004) proclaimed the onset of 'the great moderation' – a new era of stability. As Minsky argued, though, 'stability is destabilizing'. In his view, if the government is going to provide a safety net to prop up and 'validate' risky behaviour, then the other side of the coin must be *greater* oversight and regulation, not less. With rapid financial innovation, reduced regulatory oversight, and less fear of a debt deflation process, financial fragility would build until a collapse.

Money Manager Capitalism and the Crisis

Beginning in 2007, the world faced the worst economic crisis since the 1930s. References to Keynesian theory and policy became commonplace, with only truly committed free marketeers arguing against massive government spending to cushion the collapse and re-regulation to prevent future crises. All sorts of explanations were proffered for the causes of the crisis: lax regulation and oversight, rising inequality that encouraged

households to borrow to support spending, greed and irrational exuberance, and excessive global liquidity – spurred by easy money policy in the USA and by US current account deficits that flooded the world with too many dollars. While each of these explanations does capture some aspect of the crisis, none of them fully recognizes the systemic nature of the global crisis.

Unfortunately, Minsky died in 1996, but after the crash, his work enjoyed unprecedented interest, with many calling this the 'Minsky Moment' or 'Minsky Crisis'. (Cassidy 2008; Chancellor 2007; McCulley 2007; Whalen 2007) I argued above that we should not view this as a 'moment' that can be traced to recent developments. Rather, as Minsky had been arguing for nearly fifty years, what we have seen is a slow transformation of the global financial system toward what Minsky called 'money manager capitalism' that finally collapsed in 2007. Hence I call it the 'Minsky half-century' (Wray 2009).

It is essential to recognize that we have had a long series of crises in the USA and abroad, and the trend has been toward more severe and more frequent crises: muni bonds in the mid-1960s; real estate investment trusts in the early 1970s; developing country debt in the early 1980s; commercial real estate, junk bonds and the thrift crisis in the USA (with banking crises in many other nations) in the 1980s; stock market crashes in 1987 and again in 2000 with the dot-com bust; the Japanese meltdown from the early 1980s; Long Term Capital Management, the Russian default and Asian debt crises in the late 1990s; and so on. Until the current crisis, each of these was resolved (some more painfully than others – impacts were particularly severe and long-lasting in the developing world) with some combination of central bank or international institution (IMF, World Bank) intervention plus a fiscal rescue (often taking the form of US Treasury spending of last resort to prop up the US economy to maintain imports that helped to restore rest of world growth).

According to Minsky, the problem is money manager capitalism – the economic system characterized by highly leveraged funds seeking maximum returns in an environment that systematically under-prices risk (Wray 2009). There are a number of reasons for this. For example, there was the belief in the Greenspan 'put' (the Chairman would always intervene to bail out financial markets if problems developed) and the Bernanke 'great moderation' – both of which lowered perceived risk. Since the last depression and debt deflation had occurred so long ago, few market participants had any memory of it; indeed, many of those in markets did not even remember the savings and loan crisis of the 1980s! Many of the models that were used to price assets were based on a very short time horizon (five years or less; sometimes this was necessitated by the fact that the financial instruments did not exist previous to that), a period that was unusually quiescent. Further, the rise of 'shadow banks' (financial institutions that often had lower costs and less regulation) led to a competitive reduction of risk spreads (pushing interest rates on riskier assets down relative to those on safe assets). Credit ratings agencies played an important role, providing high ratings to assets that proved to be very much riskier than indicated. All of this was made worse by a general 'euphoric' belief that prices of assets (such as real estate and commodities) could only go up. Finally, there was an explosion of various types of derivatives that appeared to reduce risk by shifting it to institutions better able to absorb losses. Perhaps the best example was the use of credit default swaps that were used as insurance in case of default; but when the crisis began, it turned out that all the risk came back in the form of counterparty risk (AIG, the seller of the 'insurance', could not cover the losses). While we cannot go into all the details here, it was even worse than that because credit default swaps were also used as pure bets on failure (the bettor would win if the assets went bad), and prices of these instruments were used as indicators of the probability of default (rising credit default swap prices could induce credit raters to lower ratings, which then triggered pay-offs on the bets even as they raised borrowing costs for the debtors) (see Wray 2009).

In sum, contrary to efficient markets theory, markets generate perverse incentives for excess risk, punishing the timid with low returns (Cassidy 2009). Any money manager who tried

to swim against the stream by avoiding excessive leverage and complex and hard-to-value assets found it hard to retain clients. Those playing along were rewarded with high returns because highly leveraged funding drives up prices for the underlying assets – whether they are dot-com stocks, Las Vegas homes, or corn futures. It all works – until it doesn't. We now know from internal emails that many financial market participants knew that risk was under-priced, but adopted an 'I'll be gone, you'll be gone' strategy – take the risk, get the millions of dollars in compensation now, and retire when the whole thing collapses.

Many have accurately described the phenomenon as 'financialization' – growing debt that leverages income flows and wealth. At the 2007 peak, total debt in the US reached a record 5 times GDP (versus 3 times GDP in 1929), with most of that private debt of households and firms. From 1996 until 2007 the US private sector spent more than its income (running deficits that increased debt) every year except during the recession that followed the dot-com bust in 2000. Financial institution debt also grew spectacularly over the two decades preceding the crisis, totaling more than GDP. Exotic financial instruments exploded – outstanding credit default swaps (bets on default by households, firms, and even countries) reached over $60 trillion, and total financial derivatives (including interest rate swaps, and exchange rate swaps) reached perhaps $600 *trillion* – many times world GDP.

Some accounts blame subprime mortgages (home loans made to riskier borrowers, typically low income households) for the global financial collapse – but that is too simple. The total value of riskier mortgage loans made in the USA during the real estate boom could not have totalled more than a trillion or two dollars (big numbers but small relative to the total volume of financial instruments). The USA was not the only country that experienced a speculative boom in real estate – Ireland, Spain and some countries in eastern Europe also had them. Then there was also speculation in commodities markets – leading to the biggest boom in history, followed by the inevitable crash – that involved about a half trillion dollars of managed money (mostly US pension funds) placing bets in commodities futures markets (Wray 2008). Global stock markets also enjoyed a renewed speculative hysteria. Big banks like Goldman Sachs speculated against US state governments, as well as countries like Greece. (For example, Goldman Sachs encouraged clients to bet against the debt issued by at least 11 US states – while collecting fees from those states for helping them to place debt. A common technique was to pool risky debt into securities, sell these to investors, then 'short' the securities using credit default swaps to bet on failure. The demand for CDSs for shorting purposes would lead to credit downgrades that raised finance costs and hastened default. The most famous shorter of mortgage debt is John Paulson, whose hedge fund asked Goldman Sachs to create toxic synthetic collateralized debt obligations (CDOs) that it could bet against. According to the US Securities and Exchange Commission, Goldman allowed Paulson's firm to increase the probability of success by picking particularly risky MBSs to include in the CDOs. Goldman arranged a total of 25 such deals, named Abacus, totaling about $11 billion. Out of 500 CDOs analyzed by UBS, only two did worse than Goldman's Abacus. Just how toxic were these CDOs? Only five months after creating one of these Abacus CDOs, the ratings of 84% of the underlying mortgages had been downgraded. By betting against them, Goldman and Paulson won – Paulson pocketed $1 billion on the Abacus deals (he made a total of $5.7 billion shorting mortgage-based instruments in a span of two years) and Goldman earned fees for arranging the deals. According to the SEC Goldman's customers actually met with Paulson as the deals were assembled – but Goldman never informed them that Paulson was the shorter of the CDOs they were buying!)

On top of all this speculative fervor there was also fraud – which appears to have become normal business practice in all of the big financial institutions. It will be years, perhaps decades, before we will unravel all of the contributing factors, including the financial instruments and practices as well as the questionable activities by market players and government officials that led to the collapse. (The *Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States* (commissioned by the US

Congress and President Obama) concluded that the crisis was both foreseeable and preventable. It blamed the 'captains of finance' (heads of the biggest banks) and the 'public stewards' (officials charged with regulating the banks) for the systemic breakdown in accountability and ethics that led to the crisis. Former bank regulator William Black (who blew the whistle on Charles Keating, the convicted felon who ran Lincoln Savings, the biggest thrift to fail as a result of the 1980s crisis, and the patron of five US Senators known as the 'Keating Five') is more blunt: the biggest banks in America were run as 'control frauds' designed to enrich top management while defrauding customers and shareholders. By his reckoning, thousands of individuals committed go-to-jail fraud. Only time will tell whether they will be brought to justice.)

This much we do know: the entire financial system had evolved in a manner that made 'it' – an economic collapse and debt deflation – possible. Riskier practices had been permitted by regulators, and encouraged by rewards and incentives. Lack of oversight and prosecution led to a dramatic failure of corporate governance and risk management at most big institutions (see the *Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States*). The combination of big government and big bank interventions plus bail-outs of 'too big to fail' institutions in crisis after crisis since the 1960s let risk grow on trend. The absence of depressions allowed financial wealth to grow over the entire post-war period – including personal savings and pension funds. All of these funds needed to earn returns. As a result, the financial sector grew relative to GDP – as a percentage of value added, it grew from 10% to 20%, and its share of corporate profits quadrupled from about 10% to 40% from 1960 to 2007 (Nersisyan and Wray 2010). It simply became too large relative to the size of the economy's production and income. The crash was the market's attempt to downsize finance – just as the crash in 1929 permanently reduced the role played by finance, and allowed for the robust growth of the post-war period. Beginning in summer 2007, a series of runs on financial institutions began that would have snowballed without unprecedented intervention by governments around the world. Typically these took the form of a refusal by markets to 'refinance' banks. Recall from above that debt of financial institutions had grown tremendously, as they borrowed mostly short-term to finance positions in financial assets. Often this took the form of overnight borrowing plus very short-term commercial paper on the basis of high-quality collateral. As the crisis unfolded, borrowers had to pledge more and more collateral, and pay higher and higher interest rates to borrow. By fall of 2007, the 'haircut' (a 10% haircut means the bank can borrow 90 cents against each dollar of good collateral) was so large that many financial institutions could no longer borrow enough to finance their positions in assets – meaning they had to sell assets into a market that now feared risk. Such 'fire sales' would lead to what Irving Fisher and Minsky called a 'debt deflation'. At the same time, worried shareholders began to dump bank stocks. Without prompt rescue by governments, the 'market' would have operated in a manner that would have led to failure of most institutions. US Treasury Secretary Timothy Geithner later said that 'none of [the biggest banks] would have survived a situation in which we had let that fire try to burn itself out' and Fed Chairman Ben Bernanke said 'As a scholar of the Great Depression, I honestly believe that September and October of 2008 was the worst financial crisis in global history . . . out of maybe the 13, 13 of the most important financial institutions in the United States, 12 were at risk of failure within a period of a week or two' (*Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States*, p. 354).

It is important to include as contributing factors the erosion of New Deal institutions that had enhanced economic stability, including most importantly the creation of a high-consumption, high-employment and high-wage society. As Minsky (1986, 1996) argued, the USA emerged from the Second World War with powerful labour unions that were able to obtain good and growing wages, which fueled growth of domestic consumption out of income. According to Minsky,

debt loads were extremely low in the private sector – with debts having been paid down or wiped out by bankruptcy in the Great Depression – and with lots of safe government bonds held as assets. In combination with a strengthened government safety net (Social Security for the aged, welfare and unemployment compensation for those without jobs, the GI bill for soldiers returning home, low interest rate loans for students) this meant that consumption comprised a relatively larger part of GDP. For Minsky, consumption out of income is a very stable component – unlike investment, which is unstable. Minsky argued that investment-led growth is more unstable than growth led by a combination of consumption out of income plus government spending because the second model does not lead to worsening private sector balance sheets.

However, over the course of the past four decades, union power declined. Minsky frequently claimed that the most significant action taken during the Reagan administration was the busting of the air traffic controllers' union (which, he claimed, sent a message to all of labour). Median real wages stopped growing, consumer debt grew on trend (and then exploded after 1995), and the generosity of the safety net was reduced. Further, over the whole period, policy increasingly favoured investment and saving over consumption – with favourable tax treatment of savings and investment, and with public subsidies of business investment. Federal government also stopped growing (relative to the size of the economy) and its spending shifted away from public infrastructure investment. Inequality grew on trend, so that it actually surpassed the 1929 record inequality. President Bush even celebrated the creation of the 'ownership society' – ironically, with concentration of ownership of financial assets at the very top (Wray 2005). The only asset that was widely owned was the home, which then became the basis for a speculative real estate bubble that produced financial assets traded around the world. The global financial collapse and deep recession in the USA after 2007 then generated widespread foreclosures (13 million by 2012) – with families kicked out of their homes, owing lots of debt, and with real estate prices collapsing so that vulture hedge funds could buy up blocks of houses at pennies on the dollar. By 2010 the home ownership rate in the USA had returned to the pre-boom level.

The 1929 crash ended what Minsky and Rudolf Hilferding designated the finance capitalism stage (Wray 2009) Perhaps the global financial crisis of 2007 will prove to be the end of this stage of capitalism – the money manager phase. Of course, it is too early to even speculate on the form capitalism will take in the future. In the final section I will look at the policy response that could help to reformulate global capitalism along Minskian lines.

Minskian Policy in the Aftermath of the Collapse of Money Manager Capitalism

Minsky (1986) argued that the Great Depression represented a failure of the small-government, *laissez faire* economic model, while the New Deal promoted a Big Government/Big Bank highly successful model for financial capitalism. Following Minsky, we might say that the current crisis represents a failure of the Big Government/Neoconservative (or, outside the USA, what is called neo-liberal) model that promotes deregulation, reduced supervision and oversight, privatization, and consolidation of market power. It replaced the New Deal reforms with self-supervision of markets, with greater reliance on 'personal responsibility' as safety nets were reduced, and with monetary and fiscal policy that is biased against maintenance of full employment and adequate growth to generate rising living standards for most Americans. Even before the crisis, the USA faced record inequality, a healthcare crisis, and high rates of incarceration, among other problems facing the lower and middle classes (Wray 2000, 2005). All of these trends are important as they increase insecurity and the potential for instability, as Minsky described in one of his last published pieces (Minsky 1996).

We must return to a more sensible model, with enhanced oversight of financial institutions and with a financial structure that promotes stability

rather than speculation. We need policy that promotes rising wages for the bottom half so that borrowing is less necessary to achieve middle class living standards. We need policy that promotes employment, rather than transfer payments – or worse, incarceration – for those left behind. Monetary policy must be turned away from using rate hikes to pre-empt inflation and toward a proper role: stabilizing interest rates, direct credit controls on bank lending to prevent runaway speculation, and stronger bank supervision. (A central bank could, for example, increase margin requirements on lending to speculators, raise required down payments for bank real estate lending, and set limits on bank lending for specified purposes in a euphoric boom.)

Minsky insisted that 'the creation of new economic institutions which constrain the impact of uncertainty is necessary', arguing that the 'aim of policy is to assure that the economic prerequisites for sustaining the civil and civilized standards of an open liberal society exist. If amplified uncertainty and extremes in income maldistribution and social inequalities attenuate the economic underpinnings of democracy, then the market behavior that creates these conditions has to be constrained' (Minsky 1996, pp. 14, 15). It is time to take finance back from the clutches of Wall Street's casino.

Minsky had long called for an 'employer of last resort' program to provide jobs to those unable to find them in the private sector. In a sense this would be a counterpart to the central bank's 'lender of last resort' program. In the jobs program, government would offer a perfectly elastic supply of jobs at a basic program wage. Anyone willing to work at that wage would be guaranteed a job. Workers would be 'taken as they are' – whatever their level of education or training – and jobs would be designed for their skill level. Training would be a part of every job – to improve skills and to make workers more employable outside the program. The work would provide useful services and public infrastructure, improving living standards. While Minsky is best known for his work on financial instability, his proposal for the employer of last resort program received almost as much of his attention, especially in the 1960s and 1970s. Interested readers are referred to the growing body of work on use of job guarantee programs as part of long-term development strategy (Bhaduri 2005; Felipe et al. 2009; Hirway 2006; Minsky 1965; Mitchell and Wray 2005; Tcherneva and Wray 2007; Wray 2007). Note that this would help to achieve Minsky's goal of a high-employment economy with decent wages to finance consumption. Minsky always saw the job guarantee as a stabilizing force – and not something that is desirable only for humanitarian reasons.

The global crisis offers both grave risks as well as opportunities. Global employment and output collapsed faster than at any time since the Great Depression. Hunger and violence grew after the financial crisis – even in developed nations. The 1930s offer examples of possible responses – on the one hand, nationalism and repression (Nazi Germany), on the other a New Deal and progressive policy. From a Minskian perspective, finance played an outsized role in the run-up to the crisis, both in the developed nations, where policy promoted managed money, and in the developing nations, which were encouraged to open to international capital. Households and firms in developed nations were buried under mountains of debt even as incomes for wage earners stagnated. Developing nations were similarly swamped with external debt service commitments, while the promised benefits of Neoliberal policies often never arrived.

Minsky would probably argue that it is time to put global finance back in its proper place as a tool to achieving sustainable development, much as the USA did in the aftermath of the Great Depression. This means substantial downsizing and careful re-regulation. Government must play a bigger role, which in turn requires a new economic paradigm that recognizes the possibility of simultaneously achieving social justice, full employment, and price and currency stability through appropriate policy.

See Also

- ▶ Banking Crises
- ▶ Credit Crunch Chronology: April 2007–September 2009

- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ Euro Zone Crisis 2010

Bibliography

Bernanke, B.S. 2004. The great moderation. Speech given at the meetings of the Eastern Economics Association, Washington, DC, 20 February. Available at http://www.federalreserve.gov/Boarddocs/Speeches/2004/20040220/default/htm. Accessed 12 May 2009.

Bhaduri, A. 2005. *Development with dignity: A case for full employment*. India: National Book Trust.

Cassidy, J. 2008. The Minsky moment. *The New Yorker*, 4 February. http://www.newyorker.com/. Accessed 29 Jan 2008.

Cassidy, J. 2009. *How markets fail: The logic of economic calamities*. New York: Picador.

Chancellor, E. 2007. Ponzi Nation. Institutional Investor, 7 February.

Felipe, J., W. Mitchell, and L.R. Wray. 2009. *A reinterpretation of Pakistan's 'economic crisis' and options for policymakers*. Manuscript, Asian Development Bank.

Hirway, I. 2006. *Enhancing livelihood security through the National Employment Guarantee Act: Toward effective implementation of the Act*. The Levy Economics Institute, Working paper no. 437. http://www.levy.org/

McCulley, P. 2007. The plankton theory meets Minsky. *Global Central Bank Focus*, March. PIMCO Bonds: http://www.pimco.com/LeftNav/Featured + Market + Commentary/FF/1999–2001/FF_01_2001.htm. Accessed 8 Mar 2007.

Minsky, H.P. 1957. Central banking and money market changes. *Quarterly Journal of Economics* 71(2): 171.

Minsky, H.P. 1965. The role of employment policy. In *Poverty in America*, ed. M.S. Gordon. San Francisco: Chandler Publishing Company.

Minsky, H.P. 1975. *John Maynard Keynes*. New York: Columbia University Press.

Minsky, H.P. 1982. *Can it happen again?* Armonk: M. E. Sharpe.

Minsky, H.P. 1986. *Stabilizing an unstable economy*. New Haven/London: Yale University Press.

Minsky, H.P. 1996. *Uncertainty and the institutional structure of capitalist economies*. The Levy Economics Institute of Bard College, Working paper no. 155.

Minsky, H.P. 2008 (1987). Securitization. Levy Economics Institute of Bard College, Policy note no. 2, 12 May.

Mitchell, W.F., and L.R. Wray. 2005. In defense of employer of last resort: A response to Malcolm Sawyer. *Journal of Economic Issues* 39(1): 235–245.

Nersisyan, Y., and L.R. Wray. 2010. Transformation of the financial system: Financialization, concentration, and the shift to shadow banking. In *Minsky, crisis and development*, ed. D. Tavasci and J. Toporowski, 32–49. Basingstoke: Palgrave MacMillan.

Papadimitriou, D.B., and L.R. Wray. 1998. The economic contributions of Hyman Minsky: Varieties of capitalism and institutional reform. *Review of Political Economy* 10(2): 199–225.

Tcherneva, P.R., and L.R. Wray. 2007. *Public employment and women: The impact of Argentina's Jefes program on female heads of poor households*. Levy Economics Institute, Working paper no. 519. http://www.levyinstitute.org/publications/? docid=965

Whalen, C. 2007. *The U.S. credit crunch of 2007: A Minsky moment*. Levy Economics Institute, Public policy brief, no. 92. http://www.levy.org/

Wray, L.R. 2000. A new economic reality: Penal Keynesianism. *Challenge*, September–October, 31–59.

Wray, L.R. 2005. *The ownership society: Social security is only the beginning*. Levy Economics Institute, Public policy brief, no. 82. http://www.levy.org/

Wray, L.R. 2007. The employer of last resort programme: Could it work for developing countries? *Economic and Labour Market Papers*, 2007/5, International Labour Office, Geneva.

Wray, L.R. 2008. The commodities market bubble: Money manager capitalism and the financialization of commodities. Levy Economics Institute, Public policy brief, no. 96. http://www.levy.org/

Wray, L.R. 2009. The rise and fall of money manager capitalism: A Minskian Approach. *Cambridge Journal of Economics* 33(4): 807–828.

Minsky, Hyman (1919–1996)

Perry Mehrling

Abstract

Hyman Philip Minsky (b. 23 September 1919, d. 24 October 1996) was best known for his Financial Instability Hypothesis of the business cycle, which emphasised the dynamics of business investment finance as a recurring cause of macroeconomic instability (Minsky, Financial instability revisited: the economics of disaster. In: Reappraisal of the federal reserve discount mechanism. Board of Governors, Federal Reserve System. Reprinted as Chapter 6 in Minsky (1982), 1972; Finance and profits: the changing nature of American business cycles. In: The business cycle and public policy 1929–1980: a compendium of papers submitted to the Joint Economic Committee. Congress of the United States, 96th

Congress, 2nd Session. Government Printing Office, Washington, DC. Reprinted as Chapter 2 in Minsky (1982), 1980a). During a boom, the expansion of debt-financed investment spending causes initial 'robust' financial structures to evolve into 'fragile' financial structures, and it is this evolution that ultimately brings the expansion to an end. In the subsequent contraction, typically some fragile financial structures collapse while others are refinanced into more robust financial structures, thereby creating the preconditions for renewed expansion.

Keywords

Analytic institutionalist; Financial Instability Hypothesis; Financial innovation; Financial Keynesian; Lender of last resort

JEL Classifications

B31; B52; E44; N22; O16

Hyman Philip Minsky (b. 23 September 1919, d. 24 October 1996) was best known for his Financial Instability Hypothesis of the business cycle, which emphasised the dynamics of business investment finance as a recurring cause of macroeconomic instability (Minsky 1972, 1980a). During a boom, the expansion of debt-financed investment spending causes initial 'robust' financial structures to evolve into 'fragile' financial structures, and it is this evolution that ultimately brings the expansion to an end. In the subsequent contraction, typically some fragile financial structures collapse while others are refinanced into more robust financial structures, thereby creating the preconditions for renewed expansion.

A central reason for policy intervention in this boom–bust process, Minsky emphasised, is the ever-present danger that the contraction will get out of control and spread into a system-wide debt deflation, in which the liquidation of fragile financial structures causes a general fall in prices and profits that undermines previously robust financial structures. In this way, a normal business recession can become instead a deep and long-lasting depression, such as happened in 1929–33, when debt deflation brought down the US banking system and ushered in a depression that did not end until the Second World War, notwithstanding all the attempts of Roosevelt's New Deal (Fisher 1933). It was only wartime public spending and wartime public debt that finally created the robust financial preconditions for renewed economic expansion in the immediate post-war decades.

Minsky entered the scene as this robust growth process was getting under way, and then he watched with growing dismay as, once private debt had a chance to build up for a while, robust financial structures began to evolve into fragile financial structures. As a professor first at Brown University (1949–1958) and then at Berkeley (1958–1965), Minsky initially pursued an essentially academic interest in financial instability. His move to Washington University, St Louis (1965–1990), however, neatly coincided with the return of actual financial instability in the credit crunch of 1966. Subsequently, the opportunity to develop a consulting relationship with the fledgling Mark Twain Bank, starting in 1967, gave Minsky a ringside seat from which he watched the evolution of American banking, including the periodic financial crises that marked the stages of that evolution – 'the credit crunch of 1966, the Penn Central–Chrysler liquidity squeeze of 1969–70, the Franklin–National–REIT debacles of 1974–75, and the Hunt/Bache/Chrysler/First of Pennsylvania fiascos of 1980' (Minsky 1983). It was during these years that his mature thinking took shape.

Unfortunately, the return of financial instability coincided also with the ascendancy of a new interventionist orthodoxy in economic theory, which, extrapolating from the entirely unusual circumstances of the immediate post-war, attributed business fluctuations not to changing financial structures but rather simply to fluctuations in aggregate demand (Mehrling 2002, 2014). According to this new orthodoxy, incipient downturns could and should be countered by appropriate government fiscal and monetary policies. Government spending could maintain aggregate demand directly, and/or tax cuts and subsidies could stimulate private consumption and investment indirectly, so as to maintain aggregate income near the level of full employment.

In the short run, this policy orthodoxy achieved its stated goal, but in the longer run it acted to block the natural process of restoring robust finance, with the consequence that an increasingly fragile financial structure served as an increasing obstacle to capital investment and hence also to robust economic performance. Because of government intervention there was no debt deflation, but rather stagnation and inflation during the decade of the 1970s until finally the extraordinary tight monetary policy of 1979–82 under Paul Volcker turned the tide and created the financial precondition for renewed expansion in the decade to follow (Minsky 1986). However, the subsequent expansion was different from the immediate post-war period, because the financial preconditions were different.

What followed after Volcker was a new kind of institutional arrangement that Minsky called 'money manager capitalism', driven by a new breed of institutional investors such as pension funds, insurance companies and mutual funds. Unlike the immediate post-war period, long-term capital development of the nation was off the table, replaced instead by the pursuit of short-run financial portfolio returns. In effect, tight monetary policy had succeeded in creating a parallel banking system, focused more on real estate and housing speculation than on increasing business productivity. Minsky viewed the financial crisis of 1987 as a first crisis of this new system. He did not, however, live to see the global financial crisis of 2007–09, which the press dubbed a 'Minsky Moment'.

Intellectual Formation

Minsky got his start in economics at the University of Chicago, where he enrolled in September 1937 in the middle of the Great Depression. There is no reason to doubt Minsky's own assessment that two Chicago professors, Oscar Lange and Henry Simons, were the most significant early influences on his thought (Minsky 1985). The inspiration to study economics came from Lange, who was at that time working out a synthesis of Marx and neoclassical economics that he called market socialism (Lange 1938). Henry Simons was the source of Minsky's lifelong interest in finance, as well as the idea that the fundamental flaw of modern capitalism stemmed from its banking and financial structure (Simons 1948). Minsky took the lesson that capitalism could be stable if, first, large-scale capital investment were owned and financed publicly rather than privately and, second, smaller scale private businesses were financed with equity rather than debt.

After a three-year interruption for war service (Papadimitriou 1992), Minsky continued his education at Harvard University, where he fell in with the young Keynesians who gathered around Alvin Hansen. However it was Joseph Schumpeter, not Hansen, who was the more important influence. A 'conservative Marxist', as Minsky would later characterise his mentor, Schumpeter's earliest work on the *Theory of Economic Development* (1912) had emphasised the importance of money creation by the banking system as the crucial source of entrepreneurial finance. Banks can and do lend by creating deposits, which serve as purchasing power that entrepreneurs use to acquire the real resources they need in order to make their future plans into present realities. This mechanism, according to Schumpeter, is the source of the dynamism of capitalism, just as it is also, according to Simons, the source of capitalism's instability.

Minsky's 1954 PhD thesis 'Induced investment and business cycles' represents his attempt to insert his concerns about finance into the then-standard Hansen–Samuelson accelerator–multiplier model, which has no finance in it. Viewed in retrospect, the more fundamental contribution Minsky made in his thesis was to conceive of ordinary business firms as akin to banks, in so far as they can be seen fundamentally as cash inflow–outflow operations that confront both solvency and liquidity 'survival constraints' (1954, pp. 157–62). From this point of view, the natural accounting structure for the economic system is not the National Income and Product Accounts, which served as the empirical basis for the Hansen–Samuelson model, but rather the newer Flow of Funds accounts developed by American institutionalist Morris Copeland (1952). In effect, Minsky's mature Financial

Instability Hypothesis would build on this alternative empirical basis, though Minsky went beyond the Flow of Funds accounts to emphasise the time-dated pattern of cash commitments embedded in the structure of outstanding debts of various kinds (Minsky 1964).

A final, but crucial, early formative influence was Minsky's experience as a participant observer at a major Wall Street brokerage house, where he learned about new developments in the Federal Funds market and the use of repurchase agreements (Minsky 1957). He concluded from that experience that the goal of using monetary policy for aggregate stabilisation was probably illusory on account of the attendant financial innovation. Here was planted the seeds of the conviction that would eventually separate Minsky from post-war economic policy orthodoxy, both Keynesian and monetarist.

According to Minsky, the central bank could try to limit the supply of public bank reserves as a way of holding back expansion, but the result would only be to encourage banks to develop their own private mechanisms for economising on scarce reserves. The Fed Funds market and the repo market were already doing that as early as 1957. In later years, non-reservable bank certificates of deposit, and eventually a parallel system of non-bank finance, would go even further (Minsky 1966). With each additional step, the link between policy tools and macroeconomic outcomes became further attenuated (Minsky 1969, 1980b). And with each additional step, the divergence increased between economic policy orthodoxy and economic institutional reality, creating room for Minsky's heterodox financial instability hypothesis to gain a hearing (Mehrling 1999).

The Financial Instability Hypothesis

At the very centre of Minsky's conception of what makes a financial structure robust or fragile is the relationship between the time pattern of cash commitments and the time pattern of expected cash flows. A firm with cash flows greater than cash commitments for every future period is said to be engaged in 'hedge' finance, because the unit can meet its commitments from its own resources. Firms with so-called 'speculative' financial structures expect cash flows greater than interest payments on outstanding debt, but also anticipate the need to refinance the principal at maturity. That makes them vulnerable in the event that refinance turns out to be unexpectedly expensive or even unavailable. So-called 'Ponzi' financial structures are even more vulnerable, since their expected cash flows are insufficient even to cover interest payments, so that anticipated refinance depends on capital gains on the underlying investment as well as general financial conditions.

The core idea of the Financial Instability Hypothesis is that there is a built-in tendency for the system to shift over time from robust 'hedge' financial structures to fragile 'speculative' and 'Ponzi' financial structures. A central driver of this tendency is the apparent cheapness of short-term finance relative to long-term finance, a consequence of liquidity preference which means that wealth holders are willing to accept a lower yield on assets that are more readily (or imminently) turned into current cash. Thus it is always tempting to finance long-lived capital assets with short-term debt, planning to roll over the debt at maturity into another short-term debt. That temptation pushes firms from hedge to speculative finance.

The temptation is always there, but in the immediate aftermath of a contraction that has visibly involved the collapse of fragile financial structures, both borrowers and lenders are typically able to resist the temptation. Liquidity risk is on their minds. Thus, it is only gradually over time that robust financial structures give way to fragile financial structures, as evidence accumulates that giving in to temptation is once again a profitable strategy, for both borrowers and lenders. Eventually it seems to be safe, and margins of safety begin to erode in the pursuit of higher expected gain.

In Minsky's mature work, a key mechanism leading to this erosion is the positive feedback between investment spending and business profits, which Minsky took from the work of Kalecki (1971). When investment spending is strong, aggregate demand and hence aggregate business profits are also strong so that business cash flows exceed expectations, proving more than sufficient

to meet existing cash commitments. Thus business firms learn the lesson that their previous caution was excessive, and the road opens for a shift to more fragile finance. And of course the same mechanism works the opposite way on the way down, as lower investment leads to lower profit than expected and hence greater than expected difficulty in meeting cash commitments.

Minsky's discovery of Kalecki, probably during his sabbatical year 1969–70 at St John's College in Cambridge, England, was crucial also for shifting Minsky's view of Keynes. Back at Harvard, the Keynes that Minsky had learned was supposed to be about a liquidity preference theory of money demand, which interacted with an exogenously fixed money supply to set the rate of interest. There was nothing much in that Keynes for Minsky, with his Simons–Schumpeter vision of the integral role of elastic bank finance for business investment, and hence a commercial loan model of the (endogenous) money supply rather than Keynes' open market operations (exogenous) money model. Although Kalecki's Marxism was not the conservative type favoured by Schumpeter, it was a familiar and congenial frame to Minsky. It was through Kalecki that Minsky found his way finally to Keynes.

In his subsequent book *John Maynard Keynes* (1975), Minsky embraced Keynes in words that could apply equally to himself: 'The knowledgeable view of the operation of finance that Keynes possessed was not readily available to academic economists, and those knowledgeable about finance did not have the skeptical, aloof attitude toward capitalist enterprise necessary to understand and appreciate the basically critical attitude that permeated Keynes's work' (p. 130). In the end, Minsky came to think of his financial instability hypothesis as a completion of Keynes' work by filling in the details of the financial system, the 'logical hole' (p. 63) that Keynes left out in his own academic formulations.

Reading Keynes with his new understanding that Keynes, like himself, was always looking at the world through the lens of banking – the 'Wall Street' or 'City' view – led Minsky to formulate what he called his 'two-price theory of investment'. *Contra* the quantity theory of money, monetary conditions do not drive the price of output; but they do drive the price of capital assets. The kind of liquidity-stretching innovations that banks use to overcome central bank constraint (as in Minsky 1957) not only enable them to provide the investment finance demanded by their business clients, but also operate directly to stimulate that demand through their effect on the price of existing capital assets. Specifically, the creation of private liquidity to satisfy demand for liquidity preference lowers the premium required to hold illiquid capital assets, and hence drives up their price. Subsequently, a widening gap between the price of existing capital assets and the current output price of new capital assets provides incentive to add new capital assets to the old, which is to say incentive for debt-financed investment spending. This 'Keynesian' asset price mechanism thus offers yet another path leading from robust finance to fragile finance.

Conclusion

Minsky's Simons–Schumpeter–Kalecki–Keynes view of the world, a view very much in the tradition of indigenous American monetary thought (Mehrling 1997, 1999), put him at odds with the post-war monetary Walrasian orthodoxy of Don Patinkin, James Tobin and Franco Modigliani. Whereas orthodoxy emphasised the use of monetary policy to 'control' aggregate fluctuation, Minsky always emphasised instead the 'support' function as lender of last resort in a crisis and market maker in normal times. Regular engagement with market participants through the discount window would, Minsky thought, allow the central bank to shift the balance a bit toward hedge finance by favouring robust structures in its collateral policy. Toward that end, and inspired by the traditional practice of the Bank of England (Sayers 1936), he urged widening access to the discount window in normal times to include a broader cross-section of economic agents, not just member banks.

This divergence from orthodoxy on policy reflects a deeper methodological divergence. Whereas post-war economic orthodoxy characteristically operated within an intellectual frame of

market equilibrium, even intertemporal market equilibrium, with abstract individual agents making rational intertemporal allocation decisions, Minsky characteristically operated closer to the lived reality that actual agents confront, namely an open-ended future that is substantially uncertain (not just risky) and a present choice set that is substantially constrained by survival constraints of various kinds. Minsky's agents are not irrational, but rather more like Schumpeter's constructive entrepreneurs who imagine a possible future and then use their cash inflow–outflow interface with the economic system to acquire resources in an attempt to make that imagined future a present reality. In a world like this, it matters a lot which agents get the chance to make that attempt; it matters for the capital development of the nation.

Minsky's Financial Instability Hypothesis was designed to explain the times he was living in, not so much the post-Volcker era of money manager capitalism. At the centre of Minsky's picture is business investment finance, not household mortgage finance; bank lending, not capital market finance; and his purview is characteristically domestic not global. But the analytical apparatus he developed is more general. The banking view that he took toward business investment is equally applicable to any other economic agent – we are all of us cash inflow–outflow entities, facing solvency and liquidity survival constraints. Similarly general is Minsky's emphasis, at the level of the system as a whole, on the shifting match between the time pattern of cash commitments that is embedded in the existing structure of debt as compared to the time pattern of expected cash flow to fulfil those commitments.

See Also

- ▶ Minsky Crisis
- ▶ Post Keynesian Economics

Bibliography

Copeland, M.A. 1952. *A study of money-flows in the United States*. New York: National Bureau of Economic Research.

Fisher, I. 1933. The debt-deflation theory of great depressions. *Econometrica* 1(4): 337–357.

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy (1933–1970)*. Cambridge: Cambridge University Press.

Lange, O. 1938. *On the economic theory of socialism*. Minneapolis: University of Minnesota Press.

Mehrling, P.G. 1997. The money interest and the public interest: American monetary thought, 1920–1970. In *Harvard Economic Studies #162*. Cambridge: Harvard University Press.

Mehrling, P.G. 1999. The vision of Hyman P. Minsky. *Journal of Economic Behavior and Organization* 39(2): 129–158.

Mehrling, P.G. 2002. Don Patinkin and the origins of postwar monetary orthodoxy. *European Journal of the History of Economic Thought* 9(2): 161–185.

Mehrling, P.G. 2014. MIT and money. In *MIT and the transformation of American economics*, ed. E.R. Weintraub. History of Political Economy (supplement). Duke University Press.

Minsky, H.P. 1954. Induced investment and business cycles. Unpublished PhD Dissertation, Department of Economics, Harvard University.

Minsky, H.P. 1957. Central banking and money market changes. *Quarterly Journal of Economics,* 71(2), 171–187. Reprinted as Chapter 7 in Minsky (1982).

Minsky, H.P. 1964. Financial crisis, financial systems, and the performance of the economy. In *Private capital markets*, 173–380. Englewood Cliffs: Commission on Money and Credit Research Study. Prentice Hall.

Minsky, H.P. 1966. The evolution of American Banking: The longer view. *The Bankers' Magazine*, London, 325–329, 397–400.

Minsky, H.P. 1969. The new uses of monetary powers. *Nebraska Journal of Economics and Business*, 8. Reprinted as Chapter 8 in Minsky (1982).

Minsky, H.P. 1972. Financial instability revisited: The economics of disaster. In *Reappraisal of the Federal Reserve discount mechanism.* Board of Governors, Federal Reserve System. Reprinted as Chapter 6 in Minsky (1982).

Minsky, H.P. 1975. *John Maynard Keynes*. New York: Columbia University Press.

Minsky, H.P. 1980a. Finance and profits: the changing nature of American business cycles. In *The Business cycle and public policy 1929–1980: A compendium of papers submitted to the Joint Economic Committee*. Congress of the United States, 96th Congress, 2nd Session. Government Printing Office, Washington, DC. Reprinted as Chapter 2 in Minsky (1982).

Minsky, H.P. 1980b. The Federal Reserve: Between a rock and a hard place. *Challenge* 23(May/June): 30–36. Reprinted as Chapter 9 in Minsky (1982).

Minsky, H.P. 1982. *Can 'it' happen again? Essays on instability and finance*. Armonk: ME Sharpe.

Minsky, H.P. 1983. Institutional roots of American inflation. In *Inflation through the ages. Economic, social*

psychological and historical aspects, ed. N. Schmokler and E. Marcus, 265–277. New York: Columbia University Press.
Minsky, H.P. 1985. Beginnings. *Banca Nazionale del Lavoro Quarterly Review* 154: 211–221.
Minsky, H.P. 1986. *Stabilizing an unstable economy*, Twentieth Century Fund Report. New Haven/London: Yale University Press.
Papadimitriou, D.B. 1992. Minsky on himself. In *Essays in honor of Hyman P. Minsky*, ed. S. Fazzari and D.B. Papadimitriou, 13–26. Armonk: ME Sharpe.
Sayers, R.S. 1936. *Bank of England Operations, 1890–1914*. London: P.S. King.
Schumpeter, J.A. 1912. *The theory of economic development; an inquiry into profits, capital, credit, interest, and the business cycle*. Translated from the German by Redvers Opie, Cambridge: Harvard University Press, 1934.
Simons, H. 1948. *Economic policy for a free society.* Chicago: University of Chicago Press.

Mirabeau, Victor Riquetti, Marquis de (1715–1789)

Peter Groenewegen

Born at Perthuis, Provence, the eldest son of an aristocratic family, Mirabeau was educated by the Jesuits. He entered the army at an early age but spent much of his youth in Paris and the Versailles court in search of personal preferment. In 1737 he inherited his father's title and estate. This made possible his marriage in 1743 to Mlle de Vassan, a misalliance which produced both their famous revolutionary son, Honoré Gabriel, and prolonged lawsuits about marital property after their formal separation in 1757. Mirabeau moved to Paris in 1746, where from 1765 he held his famous salon, an activity far more successful than his management of farms and family. Long-felt literary ambitions combined with a spirit critical of government produced his first published book on provincial administration (Mirabeau 1750). In January 1757 (see Weulersse 1910, p. 20) he published the book which made him famous, gave him the title 'Friend of Mankind', put him into contact with Quesnay and converted him to Physiocracy. This well-documented conversion was followed by a large number of works in which Mirabeau either collaborated with or was heavily guided by Quesnay. The more important included new editions of *L'Ami des Hommes* (Mirabeau 1758, 1760a), a work on taxation (Mirabeau 1760b) which earned him imprisonment and brief exile from Paris for its criticism of tax farming, and the more substantial *Philosophie rurale* (Mirabeau 1763), later successfully abridged (Mirabeau 1767). Quesnay collaborated very substantially in preparing this last major work, contributing the final chapter with further explanations and manipulations of his *Tableau économique* analysis. His collaboration with Quesnay implied that Mirabeau sacrificed 'his originality, became a sectarian, and with remarkable self-abnegation, a semi-religious faith, an enthusiasm almost mystic, ... entered upon his apostleship' (Fling 1908, p. 106). Mirabeau's pre-Quesnay 'originality' can, however, be doubted. As Higgs (1931, p. 387) noted, he initially proposed to base *L'Ami des Hommes* on Cantillon's *Essai.* When that was published in 1755, Mirabeau's book developed into something more independent but was still heavily indebted to Cantillon's work. With the decline of Physiocracy in the 1770s Mirabeau's reputation waned as well. He remained in the public eye through his stormy family battles, especially his attempts to imprison his unruly son by *lettre de cachet.* He died at Argenteuil on the last day of the *ancien régime,* 13 July 1789.

Mounting evidence that Quesnay contributed most of what was original in the work published under Mirabeau's name means that his economics can now largely be regarded as a vehicle for the 'master'. His importance to economics therefore arises from the fact that he lent his well-known name, his salon and his boundless energies to propagating Quesnay's ideas. As Hecht (1958, p. 258) noted, 'From 1757, the marquis sent Quesnay even his most minor pieces, which were returned to him with annotations, commentary, criticism and corrections of style and substance'. Mirabeau's works are now largely worth studying for the light they shed on Quesnay's economics, particularly the interpretation of the *Tableau*

économique. Meek's claim (1962, p. 27) that 'just as Marx had his Engels, so Quesnay . . . had his Mirabeau' is an inappropriate analogy unless Engels is to be devalued to the role of popularizer and propagandist.

Selected Works

1750. *Mémoire concernant l'utilité des états provinciaux*. Rome (and France).
1757. *L'ami des hommes ou traité de la population*. Avignon.
1758. *L'ami des hommes ou traité de la population. Part 4 containing Questions interéssantes sur la population, l'agriculture et le commerce* (by Quesnay and Marivelt). Paris.
1760a. *L'ami des hommes ou traité de la population. Parts 5 and 6, Tableau oeconomique avec ses explications*. Paris; published as The Oeconomical Table by the Friend of Mankind, London, 1766. Extracts translated in Meek (1973).
1760b. *Théorie de l'impôt. n.p*. Reprinted, Aalen: Scientia Verlag, 1970.
1763. *Philosophie rurale ou économie générale et politique de l'agriculture.* Amsterdam/Paris. Extracts translated in Meek (1962) and (1973).
1767. *Eléments de la philosophie rurale*. The Hague/Paris.

Bibliography

Fling, F.M. 1908. *Mirabeau and the French Revolution,* vol. 1, *The Youth of Mirabeau.* New York: G.P. Putnam's Sons.
Hecht, J. 1958. La vie de François Quesnay. In *François Quesnay et la Physiocratie.* Paris: Institut National d'Etudes Démographiques.
Higgs, H. 1931. Life and work of Richard Cantillon. In Richard Cantillon, *Essay on the nature of commerce in general,* ed. Henry Higgs, London: Macmillan for the Royal Economic Society.
Meek, R.L. 1962. *The economics of physiocracy.* London: George Allen & Unwin.
Meek, R.L. 1973. *Precursors of Adam Smith 1750–1775*. London: Everyman's University Library.
Weulersse, G. 1910. *Les manuscrits économiques inédites de François Quesnay et du Marquis de Mirabeau aux Archives Nationales*. Paris.

Mirrlees, James (Born 1936)

Gareth D. Myles

Abstract

Sir James Mirrlees has been influential in several areas. With Little he contributed to development economics through 'the manual', a practical guide to the use of cost–benefit analysis. He developed the theory of optimal income taxation and in so doing introduced the concept of incentive compatibility. With Diamond he revolutionized the theory of commodity taxation. His work on the principal–agent problem characterized contracts designed to counter moral hazard. The analytical tools he pioneered have benefited every area of economics to which they have been applied.

Keywords

Asymmetric information; Border prices; Commodity taxation; Cost–benefit analysis; Endogenous growth theory; Incentive compatibility; Income taxation and optimal policies; Mirrlees, J.; Monotone likelihood ratio condition; Optimal growth; Optimum income tax; Principal–agent problem; Production efficiency lemma; Revelation principle; Shadow pricing; Single-crossing condition; Technical change; Value-added taxation; World Bank

JEL Classifications

B31

Professor Sir James Mirrlees was born in Scotland in 1936 and educated at Edinburgh University and Trinity College, Cambridge. He held academic posts at Trinity College and at Nuffield College, Oxford, and was awarded the Nobel Prize in economics in 1996 for his work on optimal income taxation and its extension to information and incentive problems in general. Mirrlees also made important contributions to growth theory, development economics and public economics.

Growth and Development

The initial work of Mirrlees focused upon technical progress in models of economic growth. Kaldor and Mirrlees (1962) assumed technical progress was embodied in new investment with the growth rate of productivity per worker operating on new machines an increasing concave function of the growth rate of investment per worker. The incorporation of externalities between different firms' investment decisions made this paper a precursor of the literature on endogenous growth theory. The problem of optimal growth in an economy subject to deterministic technical change was discussed in Mirrlees (1967). An extension to stochastic diffusion in continuous time showing that increased uncertainty would often lead to more saving rather than less was given in Mirrlees's Ph.D. dissertation and circulated in unpublished work (Mirrlees 1965). These themes were also addressed in Mirrlees (1974a).

Mirrlees contributed to development economics via the influential Little and Mirrlees (1968, 1974) handbook of project appraisal ('the manual'). The manual was a practical guide to the use of cost–benefit analysis designed to contribute to improvements in the economic conditions of developing countries. It took as its starting point the use of shadow prices to value all inputs and outputs, regardless of whether they were marketable or non-marketable, and showed how shadow prices should be determined. In particular, it emphasized the use of border prices to value inputs and outputs when the project was located in a small country. When goods were not traded, it provided methods for valuing them based on the prices of traded goods. The manual emphasized that investment finance was scarce because of the government's budget constraint, so social profits should be discounted at the internal rate of return for the marginal investment project. The manual also studied constraints upon policy choices and how these affected shadow prices.

The recommendations of the manual provided a simple but powerful methodology. Since its publication they have been subjected to much theoretical scrutiny that has generally confirmed their validity. The practical impact of the Little–Mirrlees approach can be judged from the number of donor agencies that adopted it to guide their decisions. Foremost among these was the World Bank, where cost–benefit analysis was the dominant decision-making method throughout the 1970s. Its use has steadily declined since, which is attributed by Little and Mirrlees (1994) to the changing nature of lending and the internal institutional structure of the World Bank.

Income Taxation

In a seminal paper Mirrlees (1971, p. 175) studied 'what principles should govern an optimum income tax; what such a tax schedule would look like; and what degree of inequality would remain once it was established'. Addressing this question required a model that included a motive for the redistribution of income because of endogenously generated inequality, incentive effects in labour supply and a justification for using an income tax rather than lump-sum taxation. The success of Mirlees's model was that it managed to capture all this but remained tractable and allowed the optimum tax to be characterized. No better model of income taxation has yet been proposed, although new results are still being discovered within the original framework and its specializations – see, for example, Diamond (1998), Saez (2001) and Hashimzade and Myles (2007).

The paper demonstrated that an optimum income tax leads to an allocation in which pre-tax income is increasing with ability, and that the marginal tax rate is between zero and one. Furthermore, unemployment is possible at the optimum and, when it occurs, is of the lowest-ability workers. The numerical analysis provides some of the most surprising findings. The optimal marginal rate of tax is low, at least compared with the rates applied in many countries at the time the paper was written. Furthermore, the marginal tax rate is fairly constant, so the tax function is close to being linear. These results motivated Mirlees's observation (1971, p. 207) that 'I had expected the rigorous analysis of income-taxation in the

utilitarian manner to provide an argument for high tax rates. It has not done so.'

The analysis of the model required Mirrlees to formulate and solve a series of novel theoretical problems. In doing so he developed a series of techniques that have since become standard tools of economic analysis. In the income tax problem the government must offer the workers a budget constraint along which each chooses an optimal location through utility maximization. Since the budget constraint can be nonlinear it is possible for there to be multiple optimal choices for a worker, so choice cannot be represented by a demand function. The fundamental contribution of the paper was to show how this problem could be circumvented by viewing the government as selecting an allocation (an income–consumption pair) for each worker. If every worker prefers his allocation to that of any other, then each will willingly select the allocation intended for him. This is the notion of incentive compatibility: a worker of ability level s must find that the allocation designed for someone of this ability gives at least as much utility as the allocation designed for any other ability s'. The government then conducts its optimization over the set of incentive-compatible allocations. The imposition of incentive compatibility reduces the set of feasible allocations and is responsible for the second-best nature of the optimum tax.

The paper also showed that the problem can be reduced further if workers' preferences over allocations are consistently related to ability. The restriction upon preferences introduced in Mirrlees (1971) has since become known as the single-crossing condition and implies that at every point in income–consumption space the indifference curve of a high-ability worker is flatter than that of a low-ability worker. Under single crossing, incentive compatibility requires high-ability workers to earn higher incomes and enjoy higher levels of consumption. The single-crossing condition has since found countless applications in problems involving the design of contracts for populations with agents of differing characteristics. With a continuum of consumers, it is not practical to state the incentive compatibility constraints directly. Mirlees (1971) surmounted this problem in a simple but ingenious way by showing that incentive compatibility is equivalent to utility being maximized at a worker's true skill level. The first-order condition for this optimization generates a differential equation that determines the evolution of utility as a function of ability. The differential equation can be used as a constraint on the optimization. This technique has since become known as the first-order approach to 'maximization subject to maximization'. The first-order condition is necessary but not sufficient, so there exists the possibility that the tax function arising from the optimization analysis may violate the monotonicity requirement. A direct solution to this problem is to incorporate the second-order condition into the optimization (see Ebert 1992). The 1971 income tax paper appreciated this issue, and the limitations of the first-order approach remained an issue that was addressed further in Mirlees's later work on the principal–agent problem.

That the optimum involved monotonicity implied an important observation: those with higher skills earn and consume more, so, although the government cannot directly observe skill, in equilibrium it can infer skill from income. Hence, given the optimum tax function, the announcement by a consumer of an income level is just a proxy for the direct announcement of a level of skill. This observation was later formalized in the revelation principle (Dasgupta et al. 1979; Myerson 1979) that shows it is possible to replace the income tax with an equivalent direct mechanism in which each consumer announces a skill level and, furthermore, announcing the true skill level is a dominant strategy. The revelation principle is now applied routinely in the analysis of incentive problems.

Commodity Taxation

Diamond and Mirrlees (1971a, b) revolutionized the theory of commodity taxation. The papers clarify the separation between consumer and producer prices and show that the choice of untaxed commodity is just a normalization that plays no role in determining the optimum allocation.

They were among the first to employ the emerging duality methods and used the indirect utility function to phrase the problem in terms of the after-tax consumer prices that were the natural choice variables. As well as these innovations, the commodity taxation papers contain two fundamental results. The first is the simple rule of thumb that the imposition of an optimum commodity tax system requires an equal proportionate reduction in compensated demand for all commodities. This conclusion emphasizes that the real effect of a tax system is on consumers' demands and that the effect on prices is of secondary importance. The second result, now known as the production efficiency lemma, is more surprising and of significant practical value for policy.

The production efficiency lemma states that the optimum commodity tax system results in an equilibrium that is on the frontier of the production set. There are some limitations to this result, most notably non-constant returns to scale, which imply that achieving efficiency may require some firms to be shut down, thus adversely affecting their owners' incomes. Such restrictions are clarified in Mirrlees (1972). The policy value of the lemma follows from observing that efficiency is only possible if there are no distortions in the input prices faced by producers. Input taxes should not therefore be a feature of the optimum set of commodity taxes, implying that intermediate goods should not be taxed. This observation justifies the use of value-added taxation with tax rebates available for producers who purchase intermediate goods. It also suggests that capital held by firms should not be subject to taxation, though dividends paid to consumers can and probably should be, along with their realized capital gains.

Theoretically, the production efficiency lemma is especially surprising when contrasted with the conclusions of Lipsey and Lancaster's (1956) second-best theory. The central message of Lipsey and Lancaster was that a distortion in any sector of the economy should generally be offset by introducing distortions in all other sectors. This finding had achieved great prominence at the time the Diamond–Mirrlees article was published. In contrast, the lemma states that, even when distortionary taxes and subsidies are being introduced into consumer decisions in order to redistribute real income or to finance public goods, there is no reason to distort producer decisions. This special case runs counter to the general message of Lipsey–Lancaster.

Principal–Agent

The third area to which Mirrlees made a fundamental contribution is the principal–agent problem that arises when one party wishes another to undertake an act on his or her behalf. If the act undertaken cannot be observed directly and its consequences observed only with some random error, then moral hazard can occur: the agent can attempt to hide behind the randomness to take an action which is less costly to the agent but which yields a lower expected return to the principal. Such a problem can arise in any economic relationship based on contingent contracts, for example between the owner and the manager of a firm. Mirrlees (1974b, 1975) analysed the problem facing the principal in designing a contract that provides an incentive to the agent to take the action that yields the highest expected payoff to the principal. There are considerable analytical similarities between the design of this contract and the choice of an optimum income tax. These similarities arise because the principal is choosing the contract to maximize expected payoff subject to the agent choosing an action to maximize his or her payoff. This leads again to a situation of maximization subject to maximization and its analysis via incentive compatibility.

When the agent must choose from a finite set of actions the incentive compatibility constraints can be employed directly. This is impractical for the continuous case where there would be an uncountable infinity of constraints. Consequently, it again becomes necessary to use the first-order conditions for the agent's choice problem as a constraint on the optimization of the principal. Although this had been used prior to Mirrlees's analysis of the principal–agent problem (Zeckhauser 1970), it had not been noticed that the approach might fail to generate the optimum. This possibility was made very clear in Mirrlees (1975), which provided an example where the

first-order approach failed to generate the optimum and proceeded to discuss how the problem could be overcome. The method proposed identified the possible maxima and incorporated them as constraints into the optimization. This method works but has proved unwieldy in practice, so most analyses rely on the first-order approach despite its known weaknesses. These issues were explored even further in Mirrlees (1986) and in Mirrlees and Roberts (1980).

A further issue that arises in principal–agent relationships is the conditions that guarantee the reward from the contract is monotonic: that is, the payment to the agent increases as observed output increases. If there are only two possible output levels, monotonicity arises naturally. With three possible output levels, monotonicity can easily fail (Grossman and Hart 1983). Mirrlees (1976) introduced the monotone likelihood ratio condition that is sufficient for monotonicity. This condition requires that actions that are more costly for the agent to undertake make more profitable outcomes relatively more likely. Although weaker conditions are available (Jewitt 1988), the monotone likelihood ratio condition has become another essential component of the economic theorist's toolkit. It is, of course, closely related to the single-crossing property that plays such an important role in the income tax paper.

The work of Mirrlees has contributed to the understanding of economic policy via the manual and the papers on tax policy. His work also laid the foundation for the analysis of incentive problems in the presence of asymmetric information. Taken together, incentive compatibility, the extension of the first-order approach, the single-crossing property and the monotone likelihood ratio condition provide the basic tools that no economic theorist can be without. There has not been a single area of economics in which they have not been used to great advantage.

See Also

- ▶ Incentive Compatibility
- ▶ Income Taxation and Optimal Policies
- ▶ Neoclassical Growth Theory
- ▶ Principal and Agent (i)
- ▶ Principal and Agent (ii)
- ▶ Revelation Principle

Selected Works

1962. (With N. Kaldor.) A new model of economic growth. *Review of Economic Studies* 29: 174–90.

1965. Optimal capital accumulation under uncertainty. Unpublished manuscript.

1967. Optimum growth when technology is changing. *Review of Economic Studies* 34: 95–124.

1968. (With I. Little.) *Manual of industrial project analysis in developing countries, Vol. II: Social cost–benefit analysis*. Paris: OECD.

1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.

1971a. (With P. Diamond.) Optimal taxation and public production I: Production efficiency. *American Economic Review* 61: 8–27.

1971b. (With P. Diamond.) Optimal taxation and public production II: Tax rules. *American Economic Review* 61: 261–78.

1972. On producer taxation. *Review of Economic Studies* 39: 105–11.

1974. (With I. Little.) *Project Appraisal and Planning for Developing Countries*. London: Heinemann.

1974a. Optimal allocation under uncertainty. In *Allocation Under Uncertainty*, ed. J. Drèze. London: Macmillan.

1974b. Notes on welfare economics, information and uncertainty. In *Essays in Equilibrium Behavior under Uncertainty*, ed. M. Balch, D. McFadden and S. Wu. Amsterdam: North-Holland.

1975. The theory of moral hazard and unobservable behaviour, Part I. Mimeo. Oxford: Nuffield College. Published in *Review of Economic Studies* 66 (1999): 3–21.

1976. The optimal structure of incentives and authority within an organization. *Bell Journal of Economics* 7: 105–31.

1980. (With K. Roberts.) Functions with multiple maxima. Mimeo. Oxford: Nuffield College.
1986. The theory of optimal taxation. In *Handbook of Mathematical Economics*, vol. 3, ed. K. Arrow and M. Intriligator. Amsterdam: North-Holland.
1994. (With I. Little.) The costs and benefits of analysis: project appraisal and planning twenty years on. In *Cost–benefit analysis*, ed. R. Layard and S. Glaister. Cambridge: Cambridge University Press.

Bibliography

Dasgupta, P., P. Hammond, and E. Maskin. 1979. The implementation of social choice rules: some general results in incentive compatibility. *Review of Economic Studies* 46: 185–216.
Diamond, P. 1998. Optimal income taxation: An example with a U-shaped pattern of optimal marginal tax rates. *American Economic Review* 88: 83–95.
Ebert, U. 1992. A reexamination of the optimal nonlinear income tax. *Journal of Public Economics* 49: 47–73.
Grossman, S., and O. Hart. 1983. An analysis of the principal–agent problem. *Econometrica* 51: 7–45.
Hashimzade, N., and G. Myles. 2007. The structure of the optimal income tax in the quasi-linear model. *International Journal of Economic Theory* 3: 5–33.
Jewitt, I. 1988. Justifying the first-order approach to principal–agent problems. *Econometrica* 56: 1177–1190.
Lipsey, R., and K. Lancaster. 1956. The general theory of second best. *Review of Economic Studies* 24: 11–32.
Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–73.
Saez, E. 2001. Using elasticities to derive optimal income tax rules. *Review of Economic Studies* 68: 205–229.
Zeckhauser, R. 1970. Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *Journal of Economic Theory* 2: 10–26.

Misclassification in Binary Variables

Christopher R. Bollinger

Abstract

Misclassification of binary variables is the first case of non-classical measurement error considered. Similar to the classical errors-in-variables result, misclassification of a binary regressor leads to attenuation of slope coefficient estimates in linear regression. Classical instrumental variables will not address the problem. Bounds results under a number of different sets of assumptions can be derived. When the dependent variable is binary, misclassification also leads to slope attenuation. Some identification results are available in this case.

Keywords

Binary variables; Dependent variable; Measurement error; Misclassification; Regression

JEL Classifications

C13; C25

Introduction

Measurement error has a long history in econometrics (Frisch 1934). Early work focused upon classical measurement error models, where the measurement error was modelled as additive white noise. Empirical work such as that by Poterba and Summers (1986), Mathiowetz and Duncan (1988), Stern (1989), Dwyer and Mitchell (1999), Gustman and Steinmeier (2001) and Benitez-Silva et al. (2004) documents that many binary regressors of interest are misclassified. Misclassification of binary variables is a non-classical measurement error. Let Z be the true binary variable, with $\Pr[Z = 1] = \pi$. Let X represent the observed variable. The misclassification probabilities are

$$\begin{aligned} A1 \ &: \Pr[X = 1|Z = 1] = 1 - q \\ &: \Pr[X = 1|Z = 0] = p. \\ &: p + q < 1 \end{aligned}$$

The assumption $p + q < 1$ ensures that the measurement error does not sever the relationship between X and Z or reverse the definition of the

indicated category. The relationship can be written as

$$X = Z + e.$$

In contrast to the classical EIV model $E[e|Z] = p(1-\pi) - q\pi$ and $cov(e, Z) = -(q+p)\pi(1-\pi)$ are not zero. Similarly $Cov(X, Z) = (1-p-q)\pi(1-\pi)$. The location parameter, π, of the true variable Z is not identified since $\Pr[X=1] = \mu_x = (1-p-q)\pi + p$. If the error rate is symmetric ($p = q$), π can be bounded between $[0, \mu_x]$ if $\mu_x < .5$ or $[\mu_x, 1]$ if $\mu_x > 0.5$. If the error rate is not symmetric, π cannot be bounded without further information.

Molinari (2008) considers identification regions for a general case concerning a categorical variable with multiple categories. Her results show that in the absence of any additional assumptions, the underlying probabilities for the K different categories are not identified. Assumptions about misclassification rates can lead to identified regions.

Simple Regression Model

Aigner (1973) considered a mismeasured binary regressor in a linear regression. Similar to the classical errors-in-variables result, the estimated coefficient is biased toward zero (attenuated). The simple linear model also includes the assumption that the measurement error process is independent of the regression error, and that the regression error is mean independent of the true regressor, Z. Formally:

$$\begin{aligned} &R1: y = \alpha + \beta Z + u \\ &R2: E[u|Z, X] = 0 \\ \\ &A2: \Pr[X=1|Z=1, u] = 1-q \\ &\quad : \Pr[X=1|Z=0, u] = p \\ &\quad : p + q < 1. \end{aligned}$$

It can be shown that the OLS regression of y on X estimates:

$$b = \left(\frac{\beta}{1-p-q}\right) \times \frac{(1-p-q)^2\pi(1-\pi)}{\left((1-p-q)^2\pi(1-\pi) + \pi q(1-q) + (1-\pi)p(1-p)\right)}.$$

Similar to the classical EIV bias, the attenuation is due to the second term being less than 1. Unlike the classical EIV, a linear instrumental variables approach does not identify the model. The classical IV estimator adds the assumption that a variable, W, is available and has the following properties: $Cov(W, Z) \neq 0$, $Cov(W,u) = Cov(W,e) = 0$. In the Classical EIV case, $Cov(W,X) = Cov(W,Z)$, so the ratio $\frac{Cov(y,W)}{Cov(W,X)}$ identifies the system. However, in the binary mismeasurement case the assumption that $Cov(W,e) = 0$ does not follow from the model, so $Cov(W,X) = (1-p-q)Cov(W,Z)$. Hence this ratio identifies

$$b_{IV} = \frac{Cov(y,W)}{Cov(W,X)} = \frac{\beta}{(1-p-q)}.$$

This estimate overstates the true coefficient. It is possible to use IV and OLS to establish bounds on the true parameter. Recent work by Mahajan (2006) considers identification of nonparametric models where a binary regressor is mismeasured but an instrumental type variable is available. That work is more thoroughly reviewed in Hong (2008).

Bounds for the true parameter are available in a number of contexts. Klepper (1988) considered the model when $p = q$, while Bollinger (1996) consider the more general model above. Bollinger (1996) establishes bounds for the above model as

$$b \leq \beta \leq \max\left\{\frac{\sigma_y^2}{\sigma_{xy}}\mu_x + b(1-\mu_x), b\mu_x + \frac{\sigma_y^2}{\sigma_{xy}}(1-\mu_x)\right\},$$

when $\beta > 0$. The bounds are reflected into the negatives when $\beta < 0$. One can consider 'error' in the relationship between y and Z as stemming from two sources: the regression error, u, and the

measurement error. The bounded region reflects different mixes of these two errors. Bounds for other parameters in the model (including additional regressors) are also established and Bollinger (1993) considers semi-parametric cases as well. Klepper (1988) considers bounds when additional regressors are mismeasured. Klepper (1988) considers how additional information in the form of a minimum R – *squared* for the regression in equation R1 will tighten the bounds. Bollinger (1996) considers how restrictions on (p,q) can be used to tighten the bounds. Both Bollinger (1996) and Klepper (1988) show that additional information has a dramatic impact on the bounds.

Allowing Misclassification to be Related to the Regression Error

A strong assumption in the simple model is the conditional independence of the error process from y. Kreider and Pepper (2007) relax this assumption in a model where the dependent variable is also a binary variable but correctly measured. Their results build on work by Horowitz and Manski (1995), Kreider and Hill (2009) and Manski and Pepper (2000). In general, the parameter of interest ($\beta = \Pr[y = 1|z = 1] - \Pr[y = 1|z = 1]$ here) is not identified or bounded in this case. They begin by adding the assumption

$$A3 : \Pr[X = Z] \geq v.$$

With this assumption, they show that

$$\inf_{k_1 \in \{0, \min[(1-v), \Pr[y=1,\, X=1]]\}} \left[\frac{\Pr[y=1, X=1] - k_1}{\mu_x - 2k_1 + (1-v)} - \frac{\Pr[y=1, X=0] + k_1}{(1-\mu_x) + 2k_1 - (1-v)}\right]$$
$$\times \leq \beta \leq$$
$$\sup_{k_2 \in \{0, \min[(1-v), \Pr[y=1,\, X=0]]\}} \left[\frac{\Pr[y=1, X=1] + k_2}{\mu_x + 2k_2 - (1-v)} - \frac{\Pr[y=1, X=0] - k_2}{(1-\mu_x) - 2k_2 + (1-v)}\right].$$

They then examine how two types of information affect these bounds. The first case is partial verification, where for some subset of the population, $\Pr[X = Z] \geq v_0$. This is particularly interesting when $v_0 = 1$: the case where some known subset of the population reports correctly. They also build on work by Manski and Pepper (2000) and consider the case where there is an additional monotonic instrumental variable known to affect y. Formally, they assume

$$A4 : P[y = 1|Z, W_2] \leq \Pr[y = 1|Z, W_0] \leq \Pr[y = 1|Z, W_1]$$

for some additional regressor W, where $W_1 \leq W_0 \leq W_2$. Although this assumption is not sufficient by itself to tighten the bounds, when combined with the type of lower bound in $A3$ they show that the bounds on β can be tightened significantly.

When the Misclassified Variable is the Dependent Variable

Some work has been done considering misclassification when the binary variable is the dependent variable. Typically the model is specified as having an underlying single index model:

$$P1 : \Pr[Y* = 1|X] = F\left(\underline{X}^T \underline{\beta}\right)$$

$$A5 : \Pr[Y = 1|Y* = 1, X] = 1 - q$$
$$: \Pr[Y = 1|Y* = 0, X] = p.$$

The variable Y is the observed binary variable, while Y^* is the true variable of interest. As with the models above, $p + q < 1$ is also assumed. The variables X are assumed to be correctly measured. Extending the simple location result above

$$\Pr[Y = 1|X] = (1 - p - q)F\left(\underline{X}^T\underline{\beta}\right) + p.$$

Interest lies in the derivative of these functions:

$$\begin{aligned}\left|\frac{\partial \Pr[Y = 1|X]}{\partial X}\right| &= (1 - p - q)f\left(\underline{X}^T\underline{\beta}\right)|\beta| \\ &\leq f\left(\underline{X}^T\underline{\beta}\right)|\beta| \\ &= \left|\frac{\partial \Pr[Y* = 1|X]}{\partial X}\right|.\end{aligned}$$

Estimated derivatives, if consistent for $\frac{\partial \Pr[Y=1|X]}{\partial X}$, will be attenuated but have the same sign as the true derivative. Hausman et al. (1998) examine identification in this model when the underlying model in $P1$ is a known function (for example, $F(.)$ is the cdf of the standard normal distribution). They show that identification can be achieved if $F(.)$ is a nonlinear function. The model can be estimated via MLE or NLLS. Lewbel (2000) extends this model to allow the misclassification rates to depend upon a subset of the X variables. In addition to regularity conditions on $F(.)$ and $p(X)$ and $q(X)$, he requires a variable w which is continuously distributed and does not affect the misclassification rates. He shows that the model is identified and can be estimated via non-parametric and semi-parametric approaches. Bollinger and David (1998) consider consistent estimation of this kind of model when validation data are available. They first estimate misclassification probabilities and then use quasi-MLE to estimate the model of interest.

Conclusions

There is much left to be done in the literature of misclassification of discrete variables. Molinari (2008) could be extended to consider categorical variables in a regression setting or to consider estimation of multinomial models. Only Lewbel (2000) and Bollinger and David (1998) have considered the case where the misclassification rates depend upon other variables, and both only for the case where the misclassified variable is the dependent variable.

See Also

▶ Measurement, Theory of

Bibliography

Aigner, D.J. 1973. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics* 1: 49–60.

Benitez-Silva, H., M. Buchinsky, H.M. Chan, S. Cheidvasser, and J. Rust. 2004. How large is the bias in self-reported disability? *Journal of Applied Econometrics* 19: 649–670.

Bollinger, C.R. 1993. Measurement error in a binary regressor with an application bounding the union wage differential, PhD diss., University of Wisconsin, Madison, WI.

Bollinger, C.R. 1996. Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics* 73: 387–399.

Bollinger, C.R., and M.H. David. 1998. Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association* 92(439): 827–835.

Dwyer, D., and O. Mitchell. 1999. Health problems as determinants of retirement: Are self-rated measures endogenous? *Journal of Health Economics* 18: 173–193.

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

Gustman, A.L., and T.L. Steinmeier. 2001. What people don't know about their pension and social security. In *Public policies and private pensions*, ed. W.G. Gale, J.B. Shoven, and M.J. Warshawsky. Washington, DC: Brookings Institution.

Hausman, J.A., J. Abrevaya, and F.M. Scott-Morton. 1998. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87: 239–269.

Hong, H. 2008. Measurement error models. In *The new Palgrave dictionary of economics*, 2nd ed, ed. S.N. Durlauf and L.E. Blume. Basingstoke: Palgrave Macmillan. The New Palgrave Dictionary of Economics Online, Palgrave Macmillan, 30 November 2000. http://www.dictionaryofeconomics.com/article?id=pde2008M000412; doi:10.1057/9780230226203.1076.

Horowitz, J., and C.F. Manski. 1995. Identification and robustness with contaminated and corrupted data. *Econometrica* 63: 281–302.
Klepper, S. 1988. Bounding the effects of measurement error in regressions involving dichotomous variables. *Journal of Econometrics* 37: 343–359.
Kreider, B., and S. Hill. 2009. Partially identifying treatment effects with an application to covering the uninsured. *Journal of Human Resources* 44: 409–449.
Kreider, B., and J.V. Pepper. 2007. Disability and employment: Reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association* 102(478): 432–441.
Lewbel, A. 2000. Identification of the binary choice model with misclassification. *Econometric Theory* 16(4): 603–609.
Manski, C.F., and J.V. Pepper. 2000. Monotone instrumental variables, with an application to the returns to schooling. *Econometrica* 68: 997–1010.
Mahajan, A. 2006. Identification and estimation of regression models with misclassification. *Econometrica* 74: 631–665.
Mathiowetz, N.A., and G.J. Duncan. 1988. Out of work, out of mind: Response error in retrospective reports of unemployment. *Journal of Business and Economic Statistics* 6: 221–229.
Molinari, F. 2008. Partial identification of probability distributions with misclassified data. *Journal of Econometrics* 144: 81–117.
Poterba, J.M., and L.H. Summers. 1986. Reporting errors and labor market dynamics. *Econometrica* 6: 221–229.
Stern, S. 1989. Measuring the effect of disability on labor force participation. *Journal of Human Resources* 24: 361–395.

Mises, Ludwig Edler von (1881–1973)

Murray N. Rothbard

Keywords

Austrian economics; Böhm-Bawerk, E. von; Business cycles; Cash-balance analysis; Central banking; Cuhel, F.; Currency School; Equation of exchange; Fractional reserve banking; German Historical School; Great Depression; Hayek, F. A. von; Inflation; Laissez-faire; Marginal utility of money; Mises, L. E. von; Money; Money supply; Neutrality of money; Praxeology; Price level; Purchasing power parity; Socialism; Socialist calculation debate

JEL Classifications

B31

Mises was born in Lemberg, Austria-Hungary, on 29 September 1881 and died in New York City on 18 October 1973. The son of a Viennese construction engineer for the Austrian railroads, Mises enrolled in the University of Vienna in 1900. He earned his doctorate in law and economics in 1906, after which he became a leading member of Böhm-Bawerk's famous seminar at the university. From 1913 to 1934, Mises taught as an unpaid *Privatdozent* at the University of Vienna, conducting a seminar on economic theory. From 1909 to 1934, he was an economist for the Vienna Chamber of Commerce, serving as the principal economic adviser to the Austrian government.

Disturbed at encroaching Nazi influence in Austria, Mises accepted a professorship at the Graduate Institute of International Studies in Geneva, where he taught from 1934 to 1940, after which he emigrated to New York City. Mises became a visiting professor at New York University in 1948, where he continued to teach a seminar on economic theory until he retired in 1969, spry and energetic at the age of 87.

Mises' multifaceted achievements in economic theory built upon the insights and methodology of the Menger–Böhm-Bawerk Austrian School of economics. In contrast to the Jevons and Walras branches of marginal utility theory, the Austrians engaged in a logical analysis of the action of individuals, their major focus on a step-by-step process analysis rather than on the necessarily unreal world of static general equilibrium. Furthermore, 'cause', for the Austrians, was a unilinear 'causal-genetic' flow from individual utilities and actions to price, rather than the familiar neoclassical mutual determination of mathematical functions.

Mises' first pioneering accomplishment was to extend Austrian analysis to money. In his *Theory of Money and Credit* (1912) he succeeded in integrating money into micro-theory, demonstrating how the marginal utility of money interacts with utilities of other goods and with the supply of money to determine money prices. In doing so, Mises solved 'the problem of the Austrian circle',

a formidable obstacle for any causal-genetic theorist. Since money, unlike other goods, is demanded not for its own sake but to purchase other goods in exchange, a demand to purchase and hold money must assume a pre-existing purchasing power in terms of other goods. How, then, can one explain the existence of that purchasing power; that is, of money prices? In his 'regression theorem', Mises, building on Menger's insights into the origin of money, demonstrated that the demand for money can be pushed back logically to the 'day' before the money-commodity became money, when it had purchasing power only as a commodity valuable in barter. Hence, every money must originate on the market as a valuable non-monetary commodity and cannot begin by being imposed by the state, or in an ad hoc social contract.

There were many other notable contributions in *Money and Credit.* Though superficially similar to the quantity theory of money, Mises' process analysis demonstrated the inevitable non-neutral impact of money on relative prices and incomes. Indeed, he levelled a devastating critique of such neutral-money concepts as Fisher's equation of exchange and the idea of stabilizing 'the price level'. Moreover, Mises developed a cash-balance analysis, independently of the Cambridge School and on an individualistic rather than an aggregative and holistic basis. And before Gustav Cassel, Mises set forth a purchasing-power parity theory of exchange rates under fiat money, based on a Ricardian array of goods rather than on Cassel's price-level approach (Wu 1939, pp. 115–16, 126–7, 232–5).

Money and Credit also revived the Ricardian–-Currency School insight that no quantity of the money supply can be more optimal than any other. Since money's sole function is to exchange, an increase in its quantity can only dilute the purchasing power of each money unit and can confer no social benefit. Mises concluded that fractional reserve banking, or 'circulation credit', is inflationary and distorts prices and production. He showed the ideal banking system to be 100 per cent reserves of bank notes and demand deposits to standard gold or silver. On the other hand, 8 years before C.A. Phillips (Phillips 1920), Mises showed that any individual bank is necessarily severely restricted in expanding credit, so that the abolition of central banking would go far to eliminate the problem of inflationary banking.

Finally, in analysing marginal utility, Mises incorporated the insights of the Czech Franz Cuhel (1907), a fellow member of Böhm-Bawerk's seminar, to demonstrate that marginal utility can in no sense be a measurable, mathematical quantity. Instead, it can only be a strictly ordinal subjective preference ranking; hence there can be no 'total utility' as an integral of marginal utilities. There can only be varying marginal utilities depending on the size of the 'margin', the actual unit of human choice.

Although two semesters of Böhm-Bawerk's seminar were devoted to discussing *Money and Credit,* the older Austrians resisted this new development (Mises 1978, pp. 59–60). Mises proceeded to found his own 'neo-Austrian' school, centred in his renowned biweekly private seminar at the Chamber of Commerce. Leading participants and followers included F.A. Hayek, Fritz Machlup, Gottfried von Haberler, Oskar Morgenstern, Wilhelm Ropke, Richard von Strigl, Alfred Schutz, Felix Kaufmann, Erich Voegelin, Georg Halm, Paul Rosenstein-Rodan and Lionel Robbins.

During the 1920s Mises developed (from its beginnings in *Money and Credit)* his notable theory of the business cycle, one of the few to be integrated with general micro-theory (Mises 1923–31). Formed out of the Currency School, Böhm-Bawerk's theory of capital and Wicksell's distinction between natural and loan rates of interest, Mises' 'monetary malinvestment' theory sees the boom–bust cycle as the inevitable product of inflationary credit expansion. This expansion artificially lowers interest rates and induces unsound overinvestments in higher-order capital goods, as well as underinvestment in consumer goods. Any cessation of credit expansion reveals the malinvestments and the lack of sufficient savings, and the ensuing

recession liquidates the distortions of the boom and restores a healthy economy.

Mises founded the Austrian Institute for Business Cycle Research in 1926, and his cycle theory later won attention as an explanation of the Great Depression. His most important student and follower, F.A. Hayek, who had elaborated on the theory, emigrated to the London School of Economics in 1931 and strongly influenced a rising generation of English economists. Unfortunately, most of this influence was swept away in the flush of enthusiasm for the Keynesian Revolution.

When socialism emerged after the First World War, Mises wrote a classic article (Mises 1920, 1922), demonstrating that a socialist government could not calculate economically and therefore could not organize a complex industrial economy. For two decades, socialists in Europe tried to rebut Mises' contentions, but not only had he anticipated their objections, he explicitly refuted them in the late 1940s (Mises 1949; Hoff 1949). If socialism could not calculate, and state interventionism only creates problems in the name of solving them (Mises 1929), then the only viable and truly prosperous economy is laissez-faire. In a century marked by accelerating statism and collectivism, Mises stood out among scholars as an uncompromising stalwart of laissez-faire (Mises 1927).

Austrian economists had virtually begun with defence of economic theory against the German Historical School (Menger 1883). Amidst a rising tide of logical positivism, Mises now set forth and elaborated 'praxeology', the methodology of Nassau Senior and of the Austrians (Bowley 1937). In contrast to the physical sciences, economic laws are discovered by logical deduction from self-evident axioms, such as that human beings exist and pursue goals. Praxeology develops the logical implications of the fact of individual human action (Mises 1933, 1949). Historical events are the complex resultants of many causal factors; they are not simple, homogeneous events that can, as in the positivist schema, be used to 'test' theory. Instead, prior theory must be used to explain and understand history (Mises 1957; Robbins 1932; Kirzner 1960).

In the culmination of his life-work, Mises put his methodological precepts into practice by constructing a systematic edifice of economic theory, completing the neo-Austrian integration of micro- and macroeconomics. First published in German in 1940, this monumental treatise was refined and expanded in his English-language *Human Action* (Mises 1949). Some notable features were a resurrection of Fetter's pure time-preference theory of interest; a theory of subjective costs; and a dynamic emphasis on profit-and-loss as the motive power of the economy, and on profit as a reward for successful entrepreneurial forecasting.

Even though an exile late in life, the trend of the world and of academia against him, and remaining only a visiting professor, Mises maintained his good cheer and productivity and gradually built up a new group of followers in the United States. Since his death there has been a veritable renaissance of interest in his thought and works, including the establishment of an institute in his name at Auburn University (Moss 1976; Andrews 1981; Kirzner 1982; Rothbard 1973).

Selected Works

1912. *The theory of money and credit,* 3rd English ed. Indianapolis: Liberty Classics, 1981.

1920. Economic calculation in the socialist commonwealth. In *Collectivist economic planning: Critical studies on the possibilities of socialism,* ed. F. von Hayek. London: Routledge & Sons, 1935.

1922. *Socialism: An economic and sociological analysis,* 3rd English ed. Indianapolis: Liberty Classics, 1981.

1923–31. *On the manipulation of money and credit,* ed. P.E. Greaves. Dobbs Ferry: Free Market Books, 1978.

1927. *Liberalism,* 3rd English ed. Irvington-on-Hudson: Foundation for Economic Education, 1985.

1929. *A critique of interventionism.* New Rochelle: Arlington House, 1977.

1933. *Epistemological problems of economics.* New York: New York University Press, 1981.
1940. *Nationalökonomie: Theorie des Handelns und Wirtschaftens.* Geneva: Editions Union.
1944a. *Bureaucracy.* New Haven: Yale University Press.
1944b. *Omnipotent government: The rise of the total state and total war.* Springs Mill: Libertarian Press, 1985.
1949. *Human action: A treatise on economics,* 3rd ed. Chicago: Regnery, 1966.
1957. *Theory and history: An interpretation of social and economic evolution,* 2nd ed. Auburn: Ludwig von Mises Institute, 1985.
1978. *Notes and recollections.* South Holland: Libertarian Press.

Bibliography

Andrews, J.K., ed. 1981. *Homage to Mises: The first hundred years*. Hillsdale: Hillsdale College Press.
Bien, B., ed. 1969. *The works of Ludwig von Mises*. Irvington-on-Hudson: Foundation for Economic Education.
Bowley, M. 1937. *Nassau senior and classical economics.* New York: Kelley, 1949.
Cuhel, F. 1907. *Zur Lehre von den Bedurfrissen. Theoretische Untersuchungen über das Grenzgebiet der Ökonomik und Psychologie*. Innsbruck: Wagner.
Hoff, T.J.B. 1949. *Economic calculation in the socialist society.* London: William Hodge.
Kirzner, I.M. 1960. *The economic point of view: An essay in the history of economic thought*. Princeton: Van Nostrand.
Kirzner, I.M., ed. 1982. *Method, process, and Austrian economics: Essays in honor of Ludwig von Mises*. Lexington: Lexington Books.
Menger, C. 1883. *Problems of economics and sociology,* ed. L. Schneider. Urbana: University of Illinois Press, 1963.
Mises, M. von. 1976. *My years with Ludwig von Mises*, 2nd ed. Cedar Falls: Center for Futures Education, 1984.
Moss, L., ed. 1976. *The economics of Ludwig von Mises: Toward a critical reappraisal*. Kansas City: Sheed & Ward.
Phillips, C.A. 1920. *Bank credit*. New York: Macmillan.
Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.
Rothbard, M. 1973. *The essential von Mises,* 3rd ed. Washington, DC: Ludwig von Mises Institute, 1983.
Wu, C.-Y. 1939. *An outline of international trade theories*. London: Routledge.

Misselden, Edward (*fl.* 1608–1654)

Douglas Vickers

Keywords
Balance of trade; Free trade; Market price; Misselden, E.; Rate of interest

JEL Classifications
B31

Edward Misselden, merchant-economist, held a number of appointments, including that of Deputy Governor of the Merchant Adventurers' Company at Delft, 1623–33, and representative of the Merchant Adventurers and the East India Company in various trade negotiations. His economic writings stem from his testimony before the Standing Commission on Trade appointed in 1622. His *Free Trade, or the means to make trade flourish* (1622) attributes the 'decay of trade' to the 'undervaluation of His Majesty's coin', 'the want of money', 'the excess of … consuming the commodities of foreign countries', particularly luxury goods (against which he proposed sumptuary laws), the export of bullion by the East India Company, and the decay and inadequate enforcement of regulation of the cloth trades. His proposal to remedy the shortage of money by increasing the denomination of the coin would, he acknowledged, raise general commodity prices, but this would be offset by the 'quickening of trade in every man's hand' that would result from the 'plenty of money'. Landlords and creditors could be protected by requiring that 'contracts made before the raising of monies shall be paid at the value the money went at when the contracts were made'.

Misselden's *Circle of Commerce, or the Balance of Trade* (1623), a long rejoinder to Malynes's attack on his earlier work, effectively sorts out the relation between commodity trade and the international exchange rate. 'It is not the rate of exchange, whether it be higher or lower, that maketh the price of commodities dear or

cheap . . . but it is the plenty or scarcity of commodities, their use or non-use, that maketh them rise or fall in price. ' He recognized that actual market prices may deviate from what might be thought to be intrinsic or par values.

Misselden's impressive work on the definition and computation of the balance of trade (explicitly including the earnings from re-exports, profits on fisheries, and freight income) contained an estimate for the year 1621–22, made by multiplying the five per cent customs revenue by 20 to obtain trade volume data. He pointed to the idea of the self-balancing international mechanism in his claim that 'there is a fluxus and refluxus, a flood and ebbe of the monies of Christendom traded within itself: for sometimes there is more in one part . . . less in another, as one country wanteth and another aboundeth'.

While Misselden advocated a high degree of governmental intervention in the economy, particularly in the granting of exclusive international trading privileges and the regulation of quality standards in domestic trade, he generally opposed the encouragement of monopolies. The free market analogy, which reappears frequently in his arguments, pointed also to the theory of the rate of interest: 'As it is the scarcity of money that maketh the high rates of interest, so the plenty of money will make interest low better than any statute for that purpose.' Misselden's various proposals to correct the 'decay of trade' shared the widespread concern for the state of the 'idle poor'. What was given away as charity, he proposed, should be 'orderly collected and prudently ordered for the employment of the poor'.

See Also

► Mercantilism

Selected Works

1622. *Free trade, or the means to make trade flourish*. London.
1623. *The circle of commerce, or the balance of trade*. London.

Bibliography

Letiche, J.M. 1959. *Balance of payments and economic growth*. New York: Harper.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.

Mitchell, Wesley Clair (1874–1948)

Geoffrey H. Moore

Keywords
Business cycles; History of economic thought; Mitchell, W.C.; National Bureau of Economic research

JEL Classifications
B31

Wesley C. Mitchell was born in Rushville, Illinois, on 5 August 1874 and died on 29 October 1948. Most of his professional life was spent at Columbia University (1913–19, 1922–44) and as Director of Research at the National Bureau of Economic Research in New York (1920–45).

Mitchell's principal contribution to economic theory was indirect – through the emphasis he placed throughout his working life upon the need for close interaction between the development of hypotheses and testing their conformity to fact. This emphasis was explicit both in his own work on business cycles and in the research that he promoted and guided in many fields at the National Bureau. In his final report as Director (1945) he said:

> We like to think of ourselves as helping to lay the foundations of an economics that will consist of statements warranted by evidence a competent reader may judge for himself. . . Speculative systems can be quickly excogitated precisely because they do not require the economist to collect and analyze masses of data, to test hypotheses for conformity to fact, to discard those which do not fit, to invent new ones and test them until, at long last, he has established a factually valid theory.

One of the hypotheses that Mitchell formulated has generated one of the longest continued and most widely applied scientific experiments in the field of economics. The hypothesis is that in free enterprise economies business cycles are generated by the continuous interaction of economic activities which lead or lag one another by varying intervals and which differ in amplitude of fluctuation by varying amounts. The processes have been identified and their leads and amplitudes measured on an ex ante as well as ex post basis. Historical records covering a century or more have been used to test the hypothesis. Private research institutes, governmental and international agencies in more than 30 countries have set up the statistical apparatus required to test the hypothesis, keep the information up-to-date, and derive economic forecasts from it. The information is generally summarized in the form of leading, coincident and lagging indexes.

The types of economic processes that Mitchell considered crucial to his hypothesis were largely identified in his first major work on the subject, *Business Cycles* (1913). The variables included costs, prices and profits; investment decisions and investment expenditures; employment, income and consumption expenditures; interest rates, the volume of money, credit, and bank reserves; inventories and sales. In the original and subsequent treatises he observed how economic agents reacted to changes in economic conditions and how these reactions in turn affected others. To measure leads and lags he defined business cycles in such a way that their peaks and troughs could be dated. Measures of timing, amplitude and rates of change in successive cycles were devised, and summarized across cycles to find out what patterns were typical. Patterns of change within the cycle enabled Mitchell to test whether what happened during one phase had a bearing on what happened in the next, and whether the repetitive sequences corresponded with his expectations based upon economic practices and institutions.

Among the economic processes in business cycles that Mitchell stressed was the imbalance that develops between costs and prices. As an expansion in business activity proceeds, costs of production begin to rise faster than prices. This reduces profit margins and dims the outlook for future profits. This in turn prompts cutbacks in decisions to invest, and leads to reductions in sales, output, and employment. As a recession develops, costs as well as prices increase less rapidly or are reduced, but cost reduction soon begins to exceed price reduction, enhancing prospects for profits and incentives to invest. This in turn helps to bring the recession to an end and get recovery started. Mitchell and others have looked into such questions as what kinds of costs and prices behave in this manner, why they do so, and whether or not the phenomenon is widespread. Comprehensive data bearing on the matter have only become available in recent years, but they show that the pattern has continued to emerge in market-oriented economies some 70 years after Mitchell gave it a central position in his hypothesis about the self-generating character of business cycles.

One of Mitchell's great objectives was to construct a general theory of business cycles consistent with the facts of cyclical experience that he took such pains to observe and record. His respect for the value of economic theory was demonstrated not only by this objective, but also by his long concern with the history of economic thought. For many years he taught a famous course at Columbia University called Types of Economic Theory, and lecture notes taken stenographically by students were subsequently published under that title (1949). The lectures traced the historical origins of economic theories and related their development to particular legal, political, social, and economic institutions and events. Such was his interest in theory that in 1941 Mitchell allowed the theoretical portion (Part III) of his 1913 volume, *Business Cycles,* to be re-published under the title *Business Cycles and Their Causes* (1941). This early effort to construct a dynamic theory, indeed, serves well as an interpretation of Mitchell's last work, *What Happens During Business Cycles* (1951). But the self-generating theory of business cycles, which was the hallmark of Mitchell's ideas on the subject, remained and still remains to be written.

Selected Works

1903. *A history of the Greenbacks, with special reference to the economic consequences of*

their issue: 1862–65. Chicago: University of Chicago Press.
1913. *Business cycles*. Berkeley: University of California Press.
1915. The making and using of index numbers. *Bulletin of the US Bureau of Labor Statistics* 173: 5–114.
1923. (With others) *Business cycles and unemployment*. New York: McGraw-Hill.
1927. *Business cycles: The problem and its setting*. New York: National Bureau of Economic Research.
1937. *The backward art of spending money, and other essays*. New York: McGraw-Hill.
1938. (With A.F. Burns.) *Statistical indicators of cyclical revivals,* Bulletin, vol. 69. New York: NBER.
1941. *Business cycles and their causes*. Berkeley: University of California Press.
1945. *The National Bureau's First Quarter-Century*. 25th Annual Report of the National Bureau of Economic Research. New York: NBER.
1946. (With A.F. Burns.) *Measuring business cycles*. New York: NBER.
1949. *Lecture notes on types of economic theory*. New York: Kelley.
1951. *What happens during business cycles: A progress report*. New York: NBER.

Bibliography

Burns, A.F., ed. 1952. *Wesley Clair Mitchell: The economic scientist*. New York: NBER.
Mitchell, L.S. 1953. *Two lives: The story of Wesley Clair Mitchell and myself*. New York: Simon & Schuster.

Mixed Strategy Equilibrium

Mark Walker and John Wooders

Abstract

A mixed strategy is a probability distribution one uses to randomly choose among available actions in order to avoid being predictable. In a *mixed strategy equilibrium* each player in a game is using a mixed strategy, one that is best for him against the strategies the other players are using. In laboratory experiments the behaviour of inexperienced subjects has generally been inconsistent with the theory in important respects; data obtained from contests in professional sports conforms much more closely with the theory.

Keywords

Competition; Equilibrium; Game theory; Minimax strategy; Mixed strategy equilibrium; Price dispersion; Pure strategy; Quantal response equilibrium; Reinforcement learning

JEL Classifications

C9

In many strategic situations a player's success depends upon his actions being unpredictable. Competitive sports are replete with examples. One of the simplest occurs repeatedly in soccer (football): if a kicker knows which side of the goal the goalkeeper has chosen to defend, he will kick to the opposite side; and if the goalkeeper knows to which side the kicker will direct his kick, he will choose that side to defend. In the language of game theory, this is a simple 2×2 game which has no pure strategy equilibrium.

John von Neumann's (1928) theoretical formulation and analysis of such strategic situations is generally regarded as the birth of game theory. Von Neumann introduced the concept of a *mixed strategy*: each player in our soccer example should choose his Left or Right action randomly, but according to some particular binomial process. Every *zero sum* two-person game in which each player's set of available strategies is finite must have a *value* (or *security level*) for each player, and each player must have at least one *minimax* strategy – a strategy that assures him that, no matter how his opponent plays, he will achieve at least his security level for the game, in expected value terms. In many such games the minimax strategies are pure strategies, requiring no mixing; in others, they are mixed strategies.

John Nash (1950) introduced the powerful notion of *equilibrium* in games (including

non-zero-sum games and games with an arbitrary number of players): an equilibrium is a combination of strategies (one for each player) in which each player's strategy is a *best* strategy for him against the strategies all the other players are using. An equilibrium is thus a sustainable combination of strategies, in the sense that no player has an incentive to change unilaterally to a different strategy. A *mixed- strategy equilibrium* (MSE) is one in which each player is using a mixed strategy; if a game's only equilibria are mixed, we say it is an MSE game. In two-person zero-sum games there is an equivalence between minimax and equilibrium: it is an equilibrium for each player to use a minimax strategy, and an equilibrium can consist only of minimax strategies.

An example or two will be helpful. First consider the game tic-tac-toe. There are three possible outcomes: Player A wins, Player B wins, or the game ends in a draw. Fully defining the players' possible strategies is somewhat complex, but anyone who has played the game more than a few times knows that each player has a strategy that guarantees him no worse than a draw. These are the players' respective minimax strategies and they constitute an equilibrium. Since they are *pure strategies* (requiring no mixing), tic-tac-toe is not an MSE game.

A second example is the game called 'matching pennies'. Each player places a penny either heads up or tails up; the players reveal their choices to one another simultaneously; if their choices match, Player A gives his penny to Player B, otherwise Player B gives his penny to Player A. This game has only two possible outcomes and it is obviously zero-sum. Neither of a player's pure strategies (heads or tails) ensures that he won't lose. But by choosing heads or tails randomly, each with probability one-half (for example, by 'flipping' the coin), he ensures that in expected value his payoff will be zero *no matter how his opponent plays*. This 50–50 mixture of heads and tails is thus a minimax strategy for each player, and it is an MSE of the game for each player to choose his minimax strategy.

Figure 1 provides a matrix representation of matching pennies. Player A, when choosing heads or tails, is effectively choosing one of the matrix's two rows; Player B chooses one of the columns; the cell at the resulting row-and-column intersection indicates Player A's *payoff*. Player B's payoff need not be shown, since it is the negative of Player A's (as always in a zero-sum game). Matching pennies is an example of a 2×2 game: each player has two pure strategies, and the game's matrix is therefore 2×2.

		Player B: H	Player B: T
Player A	H	–1	1
	T	1	–1

Mixed Strategy Equilibrium, Fig. 1

		Goalkeeper: L	Goalkeeper: R
Kicker	L	0.4	0.9
	R	0.8	0.3

Mixed Strategy Equilibrium, Fig. 2

Figure 2 depicts our soccer example, another 2×2 MSE game. The kicker and the goalie simultaneously choose either Left or Right; the number in the resulting cell (at the row-and-column intersection) is the probability a goal will be scored, given the players' choices. The probabilities capture the fact that for each combination of choices by kicker and goalie the outcome is still random – a goal is less likely (but not impossible) when their choices match and is more likely (while not certain) when they don't. The specific probabilities will depend upon the abilities of the specific kicker and goalie: the probabilities in Fig. 2 might represent, for example, a situation in which the kicker is more effective kicking to the left half of the goal than to the right half. For the specific game in Fig. 2 it can be shown that the kicker's minimax strategy is a 50–50 mix between Left and Right and the goalie's minimax strategy is to defend Left 3/5 of the time and Right 2/5. The

reader can easily see that the value of the game is therefore 3/5, that is, in the MSE the kicker will succeed in scoring a goal 60 per cent of the time.

Non-zero-sum games and games with more than two players often have mixed strategy equilibria as well. Important examples are decisions whether to enter a competition (such as an industry, a tournament, or an auction), 'wars of attrition' (decisions about whether and when to exit a competition), and models of price dispersion (which explain how the same good may sell at different prices), as well as many others.

How do people actually behave in strategic situations that have mixed strategy equilibria? Does the MSE provide an accurate description of people's behaviour? Virtually from the moment Nash's 1950 paper was distributed in preprint, researchers began to devise experiments in which human subjects play games that have mixed strategy equilibria. The theory has not fared well in these experiments. The behaviour observed in experiments typically departs from the MSE in two ways: participants do not generally play their strategies in the proportions dictated by the game's particular MSE probability distribution; and their choices typically exhibit negative serial correlation – a player's mixed strategy in an MSE requires that his choices be independent across multiple plays, but experimental subjects tend instead to switch from one action to another more often than chance would dictate. Experimental psychologists have reported similar 'switching too often' in many experiments designed to determine people's ability to intentionally behave randomly. The evidence suggests that humans are not very good at behaving randomly.

The results from experiments were so consistently at variance with the theory that empirical analysis of the concept of MSE became all but moribund for nearly two decades, until interest was revived by Barry O'Neill's (1987) seminal paper. O'Neill pointed out that there were features of previous experiments that subtly invalidated them as tests of the theory of mixed strategy equilibrium, and he devised a clever but simple experiment that avoided these flaws. Although James Brown and Robert Rosenthal (1990) subsequently demonstrated that the behaviour of O'Neill's subjects was still inconsistent with the theory, the correspondence between theory and observation was nevertheless closer in his experiment than in prior experiments.

Mark Walker and John Wooders (2001) were the first to use field data instead of experiments to evaluate the theory of mixed strategy equilibrium. They contended that, while the rules and mechanics of a simple MSE game may be easy to learn quickly, as required in a laboratory experiment, substantial experience is nevertheless required in order to develop an understanding of the strategic subtleties of playing even simple MSE games. In short, an MSE game may be easy to play but not easy to play *well.* This fact alone may account for much of the theory's failure in laboratory experiments.

Instead of using experiments, Walker and Wooders applied the MSE theory to data from professional tennis matches. The 'serve' in tennis can be described as a 2×2 MSE game exactly like the soccer example in Fig. 2: the server chooses which direction to serve, the receiver chooses which direction to defend, and the resulting payoff is the probability the server wins the point. Walker and Wooders obtained data from matches between the best players in the world, players who have devoted their lives to the sport and should therefore be expert in the strategic subtleties of this MSE game. Play by these world-class tennis players was found to correspond quite closely to the MSE predictions. Subsequent research by others, with data from professional tennis and soccer matches, has shown a similar correspondence between theory and observed behaviour.

Thus, the empirical evidence to date indicates that MSE is effective for explaining and predicting behaviour in strategic situations at which the competitors are experts and that it is less effective when the competitors are novices, as experimental subjects typically are. This leaves several obvious open questions. In view of the enormous disparity in expertise between world-class athletes and novice experimental subjects, how can we determine, for specific players, whether the MSE yields an appropriate prediction or explanation of their play? And when MSE is

not appropriate, what *is* a good theory of play? We clearly need a generalization of current theory, one that includes MSE, that tells us in addition when MSE is 'correct', and that explains behavior when MSE is not correct. Moreover, the need for such a theory extends beyond MSE games to the theory of games more generally.

A more general theory will likely comprise either an alternative, more general notion of equilibrium or a theory of out-of-equilibrium behaviour in which some players may, with enough experience, come to play as the equilibrium theory predicts. Recent years have seen research along both lines. Among the most promising developments are the notion of quantal response equilibrium introduced by Richard McKelvey and Thomas Palfrey (1995), the theory of level-n thinking introduced by Dale Stahl and Paul Wilson (1994), and the idea of reinforcement learning developed by Ido Erev and Alvin Roth (1998).

See Also

- ▶ Game Theory
- ▶ Game Theory in Economics, Origins of
- ▶ Nash, John Forbes (Born 1928)
- ▶ Purification
- ▶ Quantal Response Equilibria
- ▶ von Neumann, John (1903–1957)

Bibliography

Brown, J., and R. Rosenthal. 1990. Testing the minimax hypothesis: A re-examination of O'Neill's game experiment. *Econometrica* 58: 1065–1081.

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.

McKelvey, R., and T. Palfrey. 1995. Quantal response equilibria for normal-form games. *Games and Economic Behavior* 10: 6–38.

Nash, J.F. 1950. Equilibrium points in N person games. *Proceedings of the National Academy of Sciences* 36: 48–49.

O'Neill, B. 1987. Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences* 84: 2106–2109.

Stahl, D., and P. Wilson. 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization* 25: 309–327.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.

Walker, M., and J. Wooders. 2001. Minimax play at Wimbledon. *American Economic Review* 91: 1521–1538.

Mixture Models

Bruce G. Lindsay and Michael Stewart

Abstract

This article discusses statistical models involving mixture distributions. As well as being useful in identifying and describing sub-populations within a mixed population, mixture models are useful data-analytic tools, providing flexible families of distributions to fit to unusually shaped data. Theoretical advances since the mid-1970s, as well as advances in computing technology, have led to the widespread use of mixture models in ecology, machine learning, genetics, medical research, psychology, reliability and survival analysis. In particular, recent advances in non-linear time series involving mixtures have helped explain various features of financial and econometric data that more traditional models cannot capture.

Keywords

ARCH models; Autoregressive models; Bootstrap; Convex models; Density estimation; Estimation; Expectation-minimization (EM) algorithm; Finite mixture distributions; Hypothesis testing; Maximum likelihood; Mixture autogressive (MAR) models; Mixture models; Nonlinear time series; Random variables; Time series analysis

JEL Classifications

C1

Suppose that $\mathcal{F} = \{F_0 : \theta \in S\}$ is a parametric family of distributions on a sample space X, and let Q denote a probability distribution defined on the parameter space S. The distribution

$$F_Q = \int F_\theta dQ(\theta)$$

is a mixture distribution. An observation X drawn from F_Q can be thought of as being obtained in a two-step procedure: first, a random Θ is drawn from the distribution Q and then, conditional on $\Theta = \theta$, X is drawn from the distribution F_θ. Suppose we have a random sample $X_1, \ldots, X_n$ from F_Q. We can view this as a missing data problem in that the 'full data' consists of pairs $(X_1, \Theta_1), \ldots, (X_n, \Theta_n)$, with $\Theta_i \sim Q$ and $X_i|\Theta_i = \theta \sim F_\theta$, but then only the first member X_i of each pair is observed; the labels Θ_i are hidden.

If the distribution Q is discrete with a finite number k of mass points $\theta_1, \ldots, \theta_k$ then we can write

$$F_Q = \sum_{j=1}^{k} q_j F_{\theta_j},$$

where $q_j = Q\{\theta_j\}$. The distribution F_Q is called a finite mixture distribution, the distributions F_θ are the component distributions and the q_j are the component weights.

There are several reasons why mixture distributions, and in particular finite mixture distributions, are of interest. First, there are many applications where the mechanism generating the data is truly of a mixture form; we sample from a population which we know or suspect is made up of several relatively homogeneous sub-populations in each of which the data of interest have the component distributions. We may wish to draw inferences, based on such a sample, relating to certain characteristics of the component sub-populations (parameters θ_j) or the relative proportions (parameters q_j) of the population in each sub-population, or both. Even the precise number of sub-populations may be unknown to us. An example is a population of fish, where the sub-populations are the yearly spawnings. Interest may focus on the relative abundances of each spawning, an unusually low proportion possibly corresponding to unfavourable conditions one year.

Second, even when there is no a priori reason to anticipate a mixture distribution, families of mixture distributions, in particular finite mixtures, provide us with particularly flexible families of probability distributions and densities which can be used to fit to unusually (skewed, long-tailed, multimodal) shaped data which would otherwise be difficult to describe with a more conventional parametric family of densities. Also, such a fit is often comparable in flexibility to a fully nonparametric estimate but structurally simpler, and often requires less subjective input, for example in terms of choosing smoothing parameters. For example, it has been shown that the very skewed log-normal density can often by well approximated by a two-or three-component mixture of normals, each with possibly different means and variances.

Third, many problems can be recast as mixture problems. An example is the problem of estimating a decreasing density function on the positive half-line. Such a density can be expressed as a mixture of uniform distributions, and, in the nonparametric maximum likelihood estimation of mixing distributions discussed below, we see that the solution to this density estimation problem follows from the solution to the general mixture problem.

Formal interest in finite mixtures dates back to at least Karl Pearson's laborious method-of-moments fitting of a two-component normal mixture to data on physical dimensions of crabs in the late 19th century. The mathematical difficulties inherent in fitting mixtures in that time have been greatly eased with the advent of the expectation-minimization (EM) algorithm in the 1970s. This algorithm yields an iterative method for computing maximum likelihood estimates (or very accurate approximations thereof) in a general missing-data situation. As mentioned above, mixtures have a natural missing-data interpretation and so the EM algorithm, together with improved computing technology, has made the

task of fitting mixtures models to data much easier, leading to a renewal of interest in them.

Fitting Finite Mixtures Using Maximum Likelihood

The EM-algorithm generates a sequence of parameter estimates each of which is guaranteed to give a larger likelihood than its predecessor. It can be used whenever the original log-likelihood $\log f_X(x;\theta)$ is difficult to maximize over θ for given x, but $f_X(x;\theta)$ can be expressed as the marginal distribution of X in a pair (X, J) whose corresponding log-likelihood $\log f_{XJ}(x, j;\theta)$ is easier to maximize over θ for given x and j. Given a 'current estimate' θ_0, the next in the sequence θ_1 is defined as the maximizer of the EM-log-likelihood $\ell_{\mathrm{EM}}(\theta; x)$ which is defined as the conditional expectation of $\log f_{XJ}(x, J;\theta)$ over the 'missing data' J given $X = x$ computed under θ_0, that is

$$\begin{aligned}&\ell_{\mathrm{EM}}(\theta; x)\\ &= E\log f(x, J;\theta) \text{ where } J \text{ has density } f_{J|X}(j|\, x;\theta_0)\\ &= f_{XJ}(x, j;\theta_0)/f_X(x;\theta_0).\end{aligned}$$

It is guaranteed that $\log f_X(x;\theta_1) \geq \log f_X(x;\theta_0)$.

If we wish to fit a finite mixture

$$f(x;Q) = \sum_{j=1}^{k} q_j f\left(x;\theta_j\right)$$

where the number of components k is known, the EM-algorithm works in almost the same way for either one or both of the q_j's or θ_j's unknown. We regard the x_i's as the observed first members of random pairs $(X_1, J_1), \ldots, (X_n, J_n)$, but the J_i's are unobserved. We can write the full data log-likelihood as

$$\sum_{i=1}^{n}\sum_{j=1}^{k} 1\{J_i = j\}\left\{\log q_j + \log f\left(x_i;\theta_j\right)\right\}$$

(here $q_j = P\{J_i = j\}$). We now outline how to go from an initial set of estimates $q_{01}, \ldots, q_{0k}, \theta_{01}, \ldots, \theta_{0k}$ to the next in the EM-sequence $q_{11}, \ldots, q_{1k}, \theta_{11}, \ldots, \theta_{1k}$. If some of these values are known, then they of course remain unchanged. The first step is to compute the posterior probabilities

$$\pi_{j|i} = P\{J_i = j|\, X_i = x_i\}$$

computed under the q_{0j}s and

$$\theta_{0j}\text{s} = \frac{q_{0j} f\left(x_i;\theta_{0j}\right)}{\sum_{j=1}^{k} q_{0j} f\left(x_i;\theta_{0j}\right)}$$

The EM-log-likelihood is then obtained by replacing the $1\{J_i = j\}$'s in the full data log-likelihood with the $\pi_{j|i}$'s; note that the EM-log-likelihood thus obtained separates into a term involving the q_j's only and one involving the θ_j's only.

If the q_j's are unknown, we maximize

$$\sum_{j=1}^{k}\log q_j\left\{\sum_{i=1}^{n}\pi_{j|i}\right\}$$

with respect to the q_j's; this is maximized at

$$q_{1j} = n^{-1}\sum_{i=1}^{n}\pi_{j|i},$$

simply the averages of the posterior probabilities over the data:

If the θ_j's are unknown, we maximize

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\pi_{j|i}\log f\left(x_i;\theta_j\right)$$

with respect to the θ_j's. Differentiating with respect to each θ_j and setting to zero yields k weighted score equations:

$$\sum_{i=1}^{n}\pi_{j|i}\frac{\partial\log f\left(x_i;\theta_j\right)}{\partial\theta_j} = 0.$$

In many common models these are easily solved. For example, in one-parameter exponential families of the form $f(x;\theta) = e^{\theta x - K(\theta)} f_0(x)$, (for example, normal with known variance, Poisson,

and so on) let $\hat{\theta}(t)$ be that value of θ that solves $K'(\theta) = t$. Then for each j one can explicitly find the EM update as

$$\theta_{j1} = \hat{\theta}\left(\frac{\sum_{i=1}^{n} \pi_{j|i} x_i}{\sum_{i=1}^{n} \pi_{j|i}}\right),$$

a known function of a $\pi_{j|i}$-weighted average of the x_is.

Further Inferences

Once the model has been fitted, further inferences may consist of confidence intervals for, or hypothesis tests concerning, the component parameters θ_j and/or the mixing proportions q_j. When the model is correctly specified (that is, there really are k components and all the q_j's are positive), the parameter estimates behave more or less in a standard fashion: they are asymptotically normal with an estimable covariance matrix, subject to the component densities $f(x; \theta_j)$ being suitably regular. Hence confidence regions can be computed in a standard fashion, bearing in mind the restrictions on the q_j's: they are non-negative and add to 1. In addition, one should be aware that, when the weights q_j are small or the parameters θ_j for two or more groups are similar, there is a sharp loss of estimating efficiency as well as good reason to be doubtful of the accuracy of asymptotic approximations. This occurs because of the near loss of identifiability of the parameters near the boundaries of the parameter space.

Hypothesis tests are perhaps not so standard, at least not for tests concerning the q_js. If one wishes to test whether an estimate $\hat{q}_j$ is significantly different from zero, the non-negativity constraints have a significant impact, at least when it comes to using large-sample χ^2 approximations to the p-values. Since such a hypothesis constrains a parameter to be on the boundary of the parameter space, the asymptotic distribution of twice the log-likelihood ratio will be a mixture of χ^2 distributions rather than a pure χ^2, on the assumption that the model is otherwise suitably regular. In such a case, a parametric bootstrap approach can be used to obtain an approximate p-value.

An Unknown Number of Components, or Completely Unknown Q

If the number of components of a putatively finite mixture is unknown, we are essentially on the same footing as knowing absolutely nothing about Q, for reasons we now explain.

For any given data-set $x_1, \ldots, x_n$ with $d \leq n$ distinct x_i's and any pre-specified Q, no matter it be discrete or continuous, so long as the likelihoods $f(x_i; \theta)$ are bounded in θ we can find a discrete $\tilde{Q}$ with $m \leq d$ support points such that Q and $\tilde{Q}$ provide exactly the same density values at the observed data. That is, for any mixing distribution Q there is a possibly different $\tilde{Q}$ yielding a finite mixture such that Q and $\tilde{Q}$ cannot be distinguished, at least in terms of the data $x_1, \ldots, x_n$. So it suffices to restrict attention to such $\tilde{Q}$s.

An implication of this, when the likelihoods are bounded in θ, is that the maximum likelihood estimate of Q over all distributions, which we denote by $\hat{Q}$, exists and is finite with at most d (the number of distinct x_is) support points. So we never need leave the realm of finite mixtures in this setting.

This is not to say, however, that an estimate of an unknown k is readily available. The number of components in $\hat{Q}$ may be an overestimate in that some support points (respectively mixing proportions) may be so close together (small) that combining them into a single point (removing them) hardly decreases the likelihood. This and other issues related to trying to infer something about the number of components in a mixture, like hypothesis tests concerning k, are difficult problems. Some problems are still open, others have solutions that are possibly too complex to be useful.

The Nonparametric Estimate of Q

When the estimate $\hat{Q}$ discussed above exists, it is discrete with at most d support points. Hence a strategy for computing it is to try to fit a finite

mixture with d components using the EM-algorithm. In many situations this yields a sensible result. More sophisticated algorithms exist however which are related to the following gradient function characterization.

The gradient function

$$D_Q(\theta) = \sum_{i=1}^{n} \left[\frac{f(x_i;\theta)}{f(x_i;Q)} - 1 \right]$$

measures the rate of increase in the log-likelihood if we remove a small amount of weight from the mixing distribution Q and put it at the point θ. Hence, for a candidate estimate Q, if for some θ we have $D_Q(\theta) > 0$, we know that we can increase the log-likelihood by putting some weight at θ.

In light of this the following result is not surprising: if the nonparametric maximum likelihood estimate $\hat{Q}$ exists, then $D_{\hat{Q}}(\theta) \leq 0$ for all θ, and the support points of $\hat{Q}$ are included in the set of values θ where $D_{\hat{Q}}(\theta) = 0$. The fact that $D_{\hat{Q}}(\theta) > 0$ for no θ makes sense; moving mass around from $\hat{Q}$ to any other θ cannot increase the likelihood.

The nonparametric version of the mixture model falls into the class of convex models, a subject with its own independent literature. Often convex models can be written as mixture models. For example, a distribution function that is concave on the positive half-line can also be written as a nonparametric mixture of the form $\int f(x;\theta)\, dQ(\theta)$ with component density $f(x;\theta) = 1\{0 < x < \theta\}/\theta$. One can deduce that the nonparametric likelihood estimator is the least concave majorant of the empirical distribution function using the above gradient characterization. See McLachlan and Peel (2000), Titterington et al. (1985) or Lindsay (1995) for further examples and other references.

Mixtures and Nonlinear Time Series

Methods related to mixtures of distributions have in recent times enjoyed a surge in popularity in finance and econometrics, in particular in the area of time series analysis. Traditional (linear) time series models, while intuitive and tractable, are well-known to be unable to capture certain features of much financial or econometric data, including variability that changes over time and marginal distributions that can be multimodal or long-tailed.

Traditional linear time series models with Gaussian innovations have marginal and conditional distributions which are Gaussian. However, in many applications both marginal and conditional distributions can be multimodal, skewed, and fat-tailed, and exhibit other non-Gaussian features. Also, series can exhibit bursts of volatility, where the variability changes in strange ways, sometimes with some dependence on past and current values of the observable series or an unobserved underlying process of 'shocks'. In several different settings, ideas of mixtures have led to new types of models that have been quite successful at capturing many of these problematic features.

One example is the mixture of autoregressive (AR) models idea. The standard autoregressive model, where the observation at time t, Y_t, has a conditional distribution, given the past Y_{t-1}, Y_{t-2}, ... of the form

$$Y_t = \theta_0 + \sum_{\ell=1}^{L} \theta_\ell Y_{t-\ell} + sZ_t,$$

where the θ_ℓ's are fixed constants and the Z_t's are independent (often standard Gaussian) random variables. Assuming $\theta_k \neq 0$ here, the model is said to be autoregressive of order L (we abbreviate this to AR(L)). The mixture version can be represented by replacing the parameter vector $\theta = (\theta_0, \theta_1, \ldots, \theta_L, s)^T$ above at each time point t with a random version $\Theta_t = (\Theta_{0t}, \Theta_{1t}, \ldots, \Theta_{Lt}, S_t)^t$, yielding

$$Y_t = \Theta_{0t} + \sum_{\ell=1}^{L} \Theta_{\ell t} Y_{t-\ell} + S_t Z_t$$

where $P(\Theta_t = \theta^{(j)}) = q_j$, each $q_j \geq 0$ and $\sum_{j=1}^{k} q_j = 1$. For each j, we have a different

AR *regime* with corresponding parameter vector $\theta^{(j)} = \left(\theta_0^{(j)}, \ldots, \theta_L^{(j)}, s_j\right)^T$ which is chosen randomly at each time point according to the probability distribution given by the q_j's, independently of Z_t and past values of the series. All regimes need not be of the same order; an $AR(L')$ regime with $L' < L$ can be obtained by just setting $\theta_{L'+1}^{(j)} = \cdots = \theta_L^{(j)} = 0$.

This so-called mixture autoregressive (MAR) model has several appealing features. Its mathematical form means that it is relatively straightforward to derive its autocorrelation function, and indeed its stationarity properties are similarly easy to derive. An interesting point here is that it is possible to have some of the component regimes non-stationary, but, so long as their mixing proportions q_j are small enough, the overall series can still have a second-order stationarity property (see time series analysis for more details). In looser terms, we can have occasional explosive behaviour but still have a series that is well-behaved in the long run. For example, when the stock market becomes volatile we can have short bursts of heightened activity which eventually settle down. Such features cannot be captured by a single AR model.

Another feature of the MAR model is that the marginal as well as conditional distributions can change with time and be multimodal. Again, during a period of stock market volatility we might expect some sharp increases and/or decreases during these periods which may result in bi- or multimodal conditional distributions. Consider the following example (simplified version of fit to IBM data from Wong and Li 2000):

$$Y_t = \begin{cases} 0.7Y_{t-1} + 0.3Y_{t-2} + 5Z_t + & \text{withprob. } 0.55; \\ 1.7Y_{t-1} - 0.7Y_{t-2} + 5Z_t + & \text{withprob. } 0.4; \\ Y_{t-1} + 20Z_t & \text{withprob. } 0.5. \end{cases}$$

If the series has been quite volatile and Y_{t-1} and Y_{t-2} are very different, say $Y_{t-1} = 200$ and $Y_{t-2} = 300$, then the conditional distribution of Y_t would be a mixture of the form

$$Y_t \sim \begin{cases} N(230, 25) & \text{withprob. } 0.55; \\ N(130, 25) & \text{withprob. } 0.4; \\ N(200, 400) & \text{withprob. } 0.5. \end{cases}$$

However, if the series had been quite stable, say with $Y_{t-1} = 200$ and $Y_{t-2} = 201$ say, then the conditional distribution would be

$$Y_t \sim \begin{cases} N(200.3, 25) & \text{withprob. } 0.55; \\ N(199.3, 25) & \text{withprob. } 0.4; \\ N(200, 400) & \text{withprob. } 0.5. \end{cases}$$

So we still have a component for increases, a component for decreases and the same component for outliers. However, the first two components are so similar that the mixture density is markedly unimodal. This example illustrates that the MAR can capture volatility as well as a changing, possibly multimodal conditional distribution.

Estimation

The mixture structure also enables maximum-likelihood estimation of unknown parameters via the EM algorithm. We briefly outline how this would work when fitting a mixture of k AR(L) regimes, although the basic steps are the same in cases where the order of each regime can differ from component to component. As in the i.i.d. case though, the question of choosing k, the number of components of the mixture, is a difficult open problem.

We can represent the mixture in terms of an unobserved label J_t at each time point which indicates which regime applies; it is equal to j with probability q_j, $j = 1, \ldots, k$. If these were known, then the full log-likelihood of observed $(y_{L+1}, \ldots, y_n)^T$ (conditional on $y_1, \ldots, y_L$) would be

$$\ell_{\text{full}}\left(\boldsymbol{q}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}\right) = \sum_{t=L+1}^{n} \left\{ \sum_{j=1}^{k} 1\{J_t = j\} \left[\log q_j + \log f\left(y_t, \ldots, y_{t-L}; \boldsymbol{\theta}^{(j)}\right) \right] \right\},$$

where $f(\cdot\,; \boldsymbol{\theta})$ is the conditional density of Y_t given $Y_{t-1}, \ldots, Y_{t-L}$ under a single AR(L) regime. We now show how a current set of estimates $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{\theta}}^{(1)}$,

$\ldots, \widetilde{\boldsymbol{\theta}}^{(k)}$ would be updated. There are two steps, an E-step and an M-step. At the E-step the missing data is set equal to its conditional expectation, given current parameter estimates and data, which here reduce to the posterior probabilities:

$$\pi_{j|t} = P\{J_t = j | Y_t = y_t, \ldots, Y_{t-L} = y_{t-L}\}$$
$$\text{computed under current estimats}$$
$$= \frac{\tilde{q}_j f\left(y_t, \ldots, y_{t-L}; \widetilde{\boldsymbol{\theta}}^{(j)}\right)}{\sum_{t=L+1}^{n} \tilde{q}_j f\left(y_t, \ldots, y_{t-L}; \widetilde{\boldsymbol{\theta}}^{(j)}\right)}$$

The M-step consists of firstly defining the EM-log-likelihood $\ell_{\mathrm{EM}}(\boldsymbol{q}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)})$ obtained by replacing $1\{J_j = t\}$ with $\pi_{j|t}$, and then maximizing over the remaining parameters. As in the i.i.d. case, the EM-log-likelihood separates into two pieces, one involving just the q_j's, which is maximized at

$$\hat{q}_j = \frac{\sum_{t=L+1}^{n} \pi_{j|t}}{n - L}$$

and another involving the other parameters of the form

$$\sum_{j=1}^{k} \sum_{t=L+1}^{n} \pi_{j|t} \log f\left(y_t, \ldots, y_{t-L}; \boldsymbol{\theta}^{(j)}\right)$$

which when differentiated partially with respect to each $\boldsymbol{\theta}^{(j)}$ yields a separate set of *weighted likelihood* equations just as in the i.i.d. case, for example,

$$\frac{\partial}{\partial \theta_0^{(j)}} \ell_{\mathrm{EM}}\left(\boldsymbol{q}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}\right)$$
$$= \sum_{t=L+1}^{n} \pi_{j|t} \frac{\partial}{\partial \theta_0} \log f(y_t, \ldots, y_{t-L}; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}.$$

Thus, if one has a computational method to obtain the maximum likelihood estimate for a straight AR(L) model, it is possible to use the same computations on this weighted form in the M-step for the more general mixture case. Note that this method is not restricted to the Gaussian-Z_t case or a linear autoregression function.

As mentioned earlier, the autocorrelation structure of the MAR model is quite straightforward to analyse; in fact, it inherits much of the simplicity of the standard AR model. One thing that one cannot obtain using an AR or MAR model is a first-order stationary series whose square exhibits some autocorrelation, which is a key feature of certain time series models designed to capture time-varying volatility. The main breakthrough in this area was the introduction of the autoregressive conditional heteroscedastic (ARCH) model for time series errors in the early 1980s by Engle, where S_t^2, the variance of the error at time t, is allowed to depend on squares of earlier errors: if Z_t's are i.i.d. mean-zero-unit-variance errors then the series $\{\varepsilon_t\}$ given by

$$\varepsilon_t = S_t Z_t; S_t = \left(\beta_0 + \sum_{\ell=1}^{M} \beta_\ell \varepsilon_{t-\ell}^2\right)$$

is an ARCH(M)-series. One can incorporate this into a mixture setting by using the same specification for the conditional mean as in the MAR case, but allowing the errors to be generated *within each regime* by a different ARCH mechanism. Hence the full specification is

$$Y_t = \Theta_{0t} + \sum_{\ell=1}^{L} \Theta_{\ell t} Y_{t-\ell} + \varepsilon_t, \varepsilon_t = S_t Z_t;$$
$$S_t = \left(B_0 + \sum_{\ell=1}^{M} B_\ell \varepsilon_{t-\ell}^2\right),$$

where now (Θ_t, B_t) takes the value $(\theta^{(j)}, \beta^{(j)})$ with probability q_j.

The resulting MAR–ARCH model combines the extra flexibility of the MAR model with the superior modelling of volatility enjoyed by ARCH series. In addition, the ability to fit several different AR–ARCH regimes provides an aid to interpretation; as in the MAR case, we can have a different regime for each of several possible reactions at each time point, and furthermore the choices (that is, conditional distributions) can change with time. The EM-algorithm can be employed in essentially the same way as the MAR model, so long as weighted maximum likelihood estimation can be performed in the M-step

for each AR-ARCH regime (allowing the possibility of non-normal errors).

Connection to Threshold Models

There is some connection between MAR and MAR-ARCH models and another class of non-linear time series known as (self-exciting) threshold autoregressive (SETAR) models. An elementary version is

$$Y_t = \begin{cases} \theta_0^{(1)} + \theta_1^{(1)} Y_{t-1} + s_1 Z_t & \text{if } Y_{t-1} < c, \\ \theta_0^{(2)} + \theta_1^{(2)} Y_{t-1} + s_2 Z_t & \text{if } Y_{t-1} \geq c, \end{cases}$$

That is, follows one of two possible AR(1) regimes, the choice depending on whether the previous value Y_{t-1} exceeds a threshold c, in contrast to the MAR model where the choice is made independently of the earlier values of the series.

It can be shown that if the Z_t's are Gaussian then the marginal distribution of the zeroth order (where $\theta_1^{(j)} \equiv 0$) is a mixture of Gaussians, permitting multimodality.

A class of models intermediate between the SETAR models and MAR involves having several AR regimes, but the choice at each time point is partly influenced by earlier values of the series, but not in a completely deterministic way. A simple version involves replacing the thresholding rule $Y_{t-1} < c$ with $Y_{t-1} + \eta_t < c$ for an independent random variable η_t. In this case, we have a mixture of AR regimes where the mixing proportions $q_j = q_j(Y_{t-1}, c)$ depend on earlier values of the series and the threshold.

These models (MAR, MAR-ARCH, SETAR and intermediate versions) are still being fully developed, however an excellent introduction is provided in Tong (1990).

Summary

Mixture distributions, particularly finite mixtures, in general permit a great increase in flexibility of modelling without an overwhelming increase in computation difficulty, while also helping in interpretation by modelling heterogeneity in a natural way. In particular, if distributions within a certain model can be fitted by maximum likelihood, then finite mixtures of distributions from the same model can in general also by fitted by maximum likelihood using the EM-algorithm. Such finite mixtures can capture heterogeneity or other complex behaviour that single components (that is, when there is no mixture) cannot capture.

See Also

- Statistical Inference
- Testing
- Time Series Analysis

Bibliography

Lindsay, B.G. 1995. *Mixture models: Theory, geometry and applications*. NSF-CBMS regional conference series in probability and statistics, vol. 5. Hayward: Institute of Mathematical Statistics/American Statistical Association.

McLachlan, G., and D. Peel. 2000. *Finite mixture models*. New York: Wiley-Interscience.

Titterington, D.M., A.F.M. Smith, and U.E. Makov. 1985. *Statistical analysis of finite mixture distributions*. Chichester: John Wiley.

Tong, H. 1990. *Nonlinear time series: A dynamical system approach*. Oxford: Clarendon Press.

Wong, C.S., and Li, W.K. 2000. On a mixture autoregressive model. *Journal of the Royal Statistical Society* Series B 62: 95–115.

Mobile Applications, the Economics of

Timothy Bresnahan, Jason Davis, Timothy Jaconette and Pai-Ling Yin

Abstract

The huge and rapid explosion in mobile devices and apps has created a fertile ground for innovation and interesting challenges for evolution in this space. Mobile app marketplaces currently feature an extremely high supply of apps, creating intense competition to get noticed by consumers. This gives corporate developers an advantage, since they have

existing marketing infrastructure to promote their apps. In the USA, popular developers are building apps for both iOS and Android platforms, allowing the coexistence of these platforms to persist. App developer strategies vary by market: those that make mobile game apps see success most often when building multiple different game apps. However, firms that make a non-game app are most successful when they focus on continuous improvements to only one app. Commercialisation strategies are also evolving: while advertising is the most common monetisation strategy, many app developers see higher revenue yields through in-app purchases. This article explores these issues in detail.

Keywords

Android; App; App developers; Apple; App marketplace; App store; Burst campaign; Cell phone; Commercialisation; Google; Industry evolution; Innovation; iOS; Mobile; Monetisation; Multihoming; Platform; Strategy; Technology

JEL Classifications

L10; L14; L17; L100; L140; L170; M37; O3

In January 2014, mobile devices were the source of 55% of all American internet usage. This represents a major shift in the consumer technology market, as desktops and laptops previously represented the bulk of internet traffic. Based on January 2014 figures, 55% of Americans over the age of 18 own smartphones and 42% own tablet computers, such as the iPad (O'Toole 2014). Additionally, in 2013, smartphones represented 65.33% of all smart connected devices sold worldwide. In the same year, portable PCs represented only 11.59% of all smart connected devices, desktop PCs represented 8.89% and the tablet plus 2-in-1 devices represented 14.19% of devices sold (IDC 2014). The mobile applications (app) industry is ripe for economic analysis due to the shifting attention of consumer internet usage. This article seeks to address the supply side of mobile platform and application strategies, with specific attention paid to the software operating systems that host applications, the applications themselves and the mobile software application developers.

Consumer mobile phone offerings are rapidly transforming from feature phones used to place calls to smartphones that serve as portable computing devices. The current mobile phone software ecosystem is dependent upon two major platforms that control software access through centralised marketplace ecosystems. Software stores of the past that sold CD-ROMs and internet downloads have been replaced by mobile application stores managed by Android and iOS, the dominant software platforms. While the American app market features software downloaded through either Google Play or iTunes, other countries such as China play host to a variety of third-party mobile application stores.

These new marketplaces and app ecosystems provide fertile ground for examining classic economic and strategic questions about innovation and industry evolution. One predominant direction of analysis is developer-centric. What strategies are developers using to make money with apps? Who is building all the apps? What is the industrial organisation of these markets, and how will the markets evolve in the future? Another line of inquiry examines platform economics, where investigators research the competition between various mobile app platforms and the network effects implicit in the expansion of certain platforms. Additionally, platforms have created economic growth opportunities despite a recession.

Current Problems in App Markets

App demand at present is minuscule compared to the total population of apps available for download. On average, a consumer's phone contains 33 apps. However, only 12 apps are typically used within a 30 day period. Ultimately, only eight apps on the phone tend to be paid; the remaining 25 tend to be free downloads (Our Mobile Planet 2013). However, the extremely low entry costs afforded by a platform structure has led numerous entrepreneurs to build an overwhelming supply of

apps on each platform. Thus, at the time of publication, we note that both platforms currently host nearly 1.5 million apps each (AppBrain 2014; Pocket Gamer 2014).

With so many apps and limited demand for an individual app, the marketplaces face a congestion problem. The main app discovery mechanism provided by the platforms is a system of top ranking lists. Only a few hundred apps out of millions can be featured on these top ranking lists at any one time. Thus, incredibly popular apps gain users quickly. However, firms without a popular app have difficulty breaking into the top rankings. 'It appears that the high costs of finding customers for new firms, some of which can be attributed to the considerable difficulty of matching buyer to seller on the platform-supplied app stores are limiting the role of entrepreneurship' (Bresnahan et al. 2014b). The app search process thus far has not solved the difficult problem of matching apps to their highest valued consumers.

The major app developers are becoming larger. App creation is moving toward becoming a winner-take-all arena, which has the ability to stifle the profitability of innovation as clear winners are established. Across both platforms, the highest ranked 20 apps make up about 80% of app use. Most of these top 20 apps serve different functions and do not compete with each other (Bresnahan et al. 2014a). This means we see clear winners at the moment for certain highly sought-after mobile functions, such as social networking or mapping. Additionally, these common functions comprise a majority of mobile use time.

Established companies tend to be the winners in a congested app marketplace. Since existing commercial entities often have a healthy base of customers, they are able to deploy marketing strategies quickly and cheaply. For example, an airline can leverage owned assets such as a website, ticket counter signage and email blasts to promote a corporate app. Many thought that the mobile revolution would shepherd a boom in entrepreneurship, allowing new companies to bloom. This has definitely happened – Snapchat, Rovio and Uber all leverage the unique features of a smartphone to craft their budding business models. However, existing companies have a very strong foothold in the minds of consumers and fulfil a sizable portion of app demand. This makes it even tougher for a breakout startup to launch into the top rankings and succeed with an app. Not only do well-funded emergent developers need to fight with over two million other apps for a spot on a top list where their apps can be shown, they also need to jockey past the existing consumer products and services companies that can leverage huge, multifaceted marketing apparatuses to promote their apps. 'In this setting, the potential for disruptive entrepreneurial innovation is diminished' (Bresnahan et al. 2014a).

When it comes to independent developers, the matching problem, coupled with congestion in mobile app stores, leads to large user acquisition costs. Marketing and app discovery are also now central to the strategic choices that app-related companies face. Firms originally hoped to be featured in the app store and gain users through organic discovery. As the app marketplace exploded, native app discovery was no longer an option for a firm launching a new app. Breaking into the top rankings was virtually impossible without either existing marketing assets or a paid promotion strategy. Thus, mobile advertising of apps became important.

One early firm strategy was the *burst campaign.* Apps would seek to acquire a number of users through various advertising networks all at the same time. If apps were able to acquire thousands of users on the same day at once, they could break into the top rankings of the mobile app stores. Once an app was displayed on the app store, or burst into the rankings, more users would download the app and use would snowball. Thus, companies and products emerged to sell inexpensive methods of acquiring users via banner advertising and other paid promotion strategies. Apps began to buy users in order to game the system of ranking charts.

New app service firms, such as TapJoy, AirPush and AppGratis, entered the marketplace. The whole purpose of this new generation of firms has been to match users with apps. Their strategies have varied; however, they provide a novel revenue model, since their success is not dependent upon fighting with the millions of apps available

in the app store. Rather, these firms live and die based upon changes to app store rules that sometimes prevent their distribution methods from permeating marketplaces, due to rule changes that limit the actions that certain software development kits can offer to apps.

Platform Competition and Multi-Homing

The smartphone industry is currently led by two major platforms in the US market. Google's Android operating system features a more open software build with greater flexibility for developers and a lower price point possibility for consumers. Apple's iOS operating system is a far more tightly controlled platform with premium luxury pricing.

At the moment, both platforms in the US smartphone market exist in near equilibrium, as seen in Fig. 1. Also, the minor platforms from RIM, Blackberry and Microsoft have been largely marginalised to irrelevance. As one can observe, both smartphone sales market share and smartphone installed base share remain neck-in-neck at time of publication. Table 1 compares the similarities and differences between platforms.

Three conditions exist to explain the current market equilibrium. '(1) The distribution of app attractiveness to consumers is skewed, with a small minority of apps drawing the vast majority of consumer demand. (2) Apps which are highly demanded on one platform tend also to be highly demanded on the other platform. These highly demanded apps have a strong tendency to multihome, writing for both platforms' (Bresnahan et al. 2014b). As popular apps are widely available on both platforms, little reason exists for users to switch suddenly from one platform to another. People who use an iPhone largely need the same apps as those who use an Android phone. Thus, evidence suggests that a major shock would be required for the market to tip.

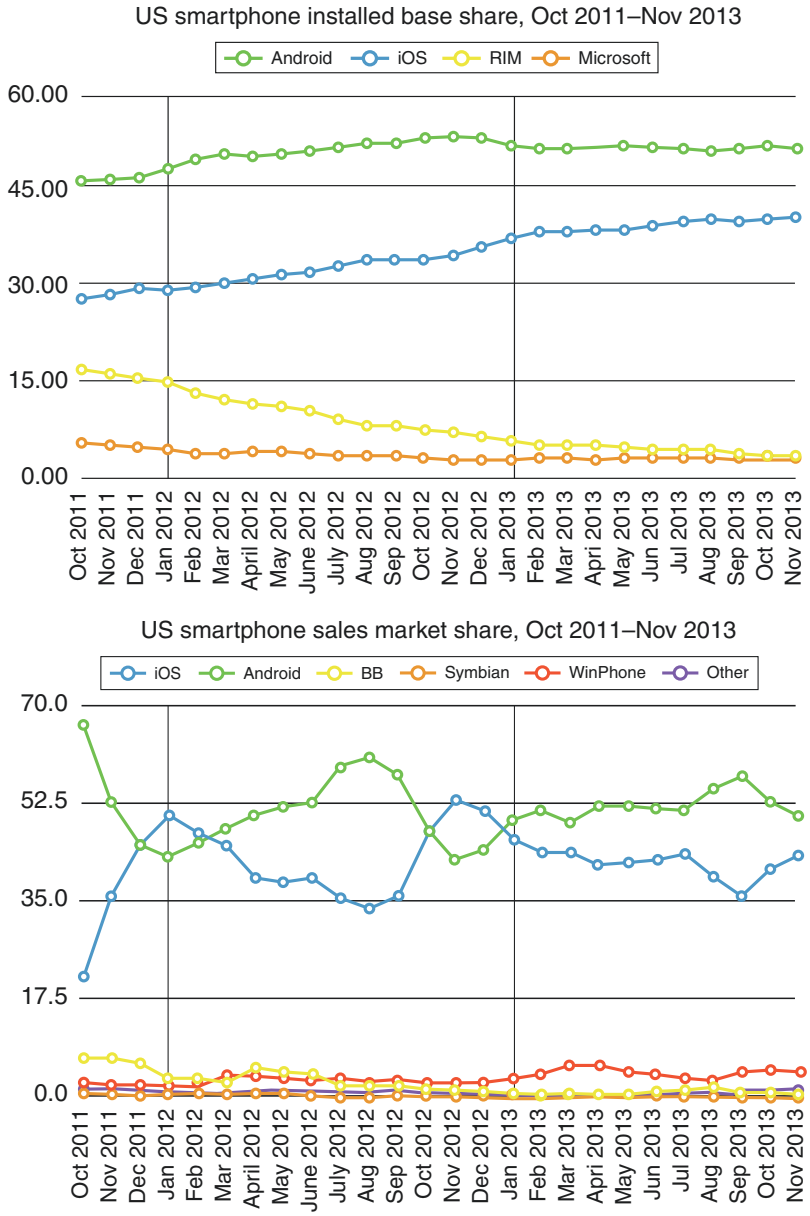
The difference between pre-existing companies, such as United Airlines, and mobile-only firms, such as Rovio, extends to multihoming behaviour. The term *multihoming* refers to the tendency of an app developer to create versions that exist on multiple platforms. A pre-existing company with an Android app has a 65% probability of building an app on both Android and iOS. A pre-existing company with an iOS app has a 58% probability of building an app on both platforms. In contrast, mobile startups with an Android app have a 23% chance of also developing an app for both platforms, and those with an iOS app have a 27% chance of developing an app for both platforms (Bresnahan et al. 2014a). Note that it is common for a mobile-only startup to launch on one platform first, then port to the other in due time. Some startups never port and maintain a business dedicated to serving the needs of one platform's users. See Fig. 2 for an overview of multihoming patterns by platform.

The world smartphone market is vastly different from the equilibrium situation in the USA. Since the cost of an iPhone can be equivalent to a month's salary or more in other countries, cheaper Android holds a tremendous lead in terms of user base around the world. In the second quarter of 2014, Android enjoyed an 84.7% worldwide smartphone market share, while iOS held an 11.7% worldwide smartphone market share (US Smartphone Installed Base Share 2014).

Firm Strategies

Many entrepreneurial firms are faced with a strategic choice of app type to develop and number of apps to build. Different app markets require different strategies. A game firm is more likely to be successful when building multiple apps. Multiple game versions allow the firm to innovate and test to see what game appeals most to customers. However, a non-game firm such as Instagram or Uber is more likely to be successful with a single app that is constantly updated and well-supported (Davis et al. 2014).

The likelihood of building a 'killer app', defined as one that appears in the top 300 apps ranked in the app store, is increased with more app updates. Furthermore, increasing the time interval between the release of new versions allows for a positive and significant impact on



Mobile Applications, the Economics of, Fig. 1 Platform growth (source: US Smartphone Installed Base Share 2014)

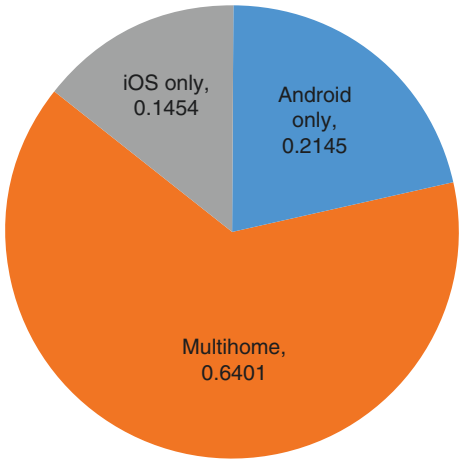
creating a 'killer app'. The simultaneous release of apps also creates a positive and significant impact on the probability of becoming a killer-app game developer.

Davis et al. (2014) 'find robust evidence that experimental strategies emphasizing many highly varied products that are released concurrently are better suited to markets that resemble more established markets and/or markets with demand for variety (games). Cumulative iteration strategies emphasizing multiple sequential versions produced over substantial time periods are better suited to nascent markets and/or markets with unit demand (non-games). In markets where well-developed innovation capabilities, knowledge of customer preferences, and technological development tools can be borrowed from more established markets and/or markets with demand for variety, experimental strategies are more suitable than cumulative strategies'.

Mobile Applications, the Economics of, Table 1 Platform characteristics – asymmetries between platforms

iOS and iTunes store	Android and Google
Early: more devices in use	Now: caught up in total devices in use
Tablets do not support flash	Tablets support flash
Richer users	Users may not buy as many in-app purchases
More restrictions on developers	Develop anywhere
Limited range of devices	Fragmentation
'Managed' change from year to year: porting an app to the newest iPhone/iPad devices from older ones is usually simple	Changes in environment from year to year, e.g. substantial UI changes
Always similar OS versions	Different hardware manufacturers use different OS versions
Distribution restricted to iTunes stores	Open, multiple distribution channels
Premium product, higher price points	Phones available for less than $50 without a contract

Source: Bresnahan et al. 2014a



Mobile Applications, the Economics of, Fig. 2 Platform choice (source: Bresnahan et al. 2014a)

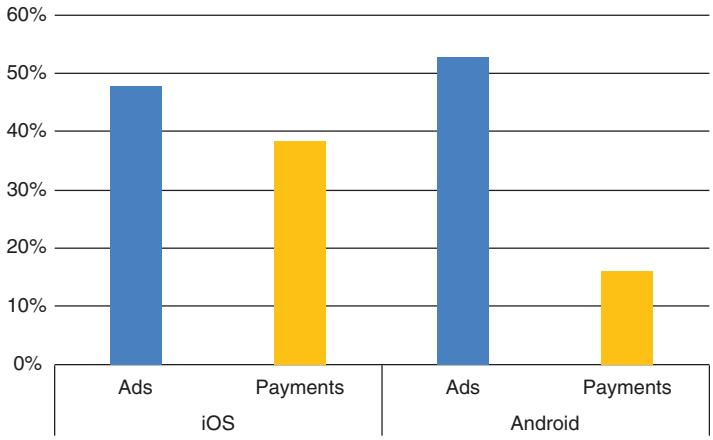
Monetisation

The main monetisation model for apps has shifted from charging 99 cents for an app to enabling in-app purchases and making the bulk of revenue from high-volume customers who spend excessively on virtual goods. Apps purchase various forms of advertising, seeking to make a profit from the difference between the cost to acquire a user and the average user's lifetime virtual goods spend within the app.

As advertisers became more sophisticated and advertising networks developed more detailed and specific user targeting mechanisms, firm strategies shifted toward complicated targeted advertising strategies. Once an app firm learned the traits of an individual user who is likely to spend a noteworthy sum of money in-app, that firm would be willing to spend more than the average advertising cost to market an app to that specific user. Major internet firms seeking to monetise existing products, such as Facebook and Twitter, jumped on this opportunity to make money from their apps by offering relevant targeted advertisements to users. Thus, app installations turned into the new form of monetising the transition of a firm's advertising-based products to mobile platforms. As top lists remained extremely difficult to break into due to the extraordinarily high bar set by top apps such as Google Maps and Facebook, firms learned to thrive by acquiring and monetising specific users.

While some firms purchase advertising to boost user counts so that they can make money from in-app purchasing, other firms rely on advertising as a revenue strategy. Some firms pursue both strategies at once. Figure 3 reveals firms' ad or payment strategies based upon survey data collected by the Stanford Mobile Innovation Group. Researchers have observed that advertising is a much more common monetisation method across both platforms than in-app purchases. This trend is logical, as banner ads can quickly and easily be incorporated into any app. No matter the category or design, the developer can easily insert an advertisement. The app does not need to be changed and the fundamental mechanics of a non-monetised app are not altered based on an advertising framework.

However, the profitability of an app that monetises solely through advertising is often limited. Average mobile banner ads are quite small and often fall victim to erroneous clicks from large thumbs, and thus do not sell for expensive rates. Over the past few years, many developers,



Mobile Applications, the Economics of, Fig. 3 In-app purchases and ads in free apps (source: Bresnahan et al. 2014a)

including Facebook and Google, have made tremendous strides in mobile app advertising innovation. That being said, cash payments from users currently trump advertising in terms of profit maximisation per app. The average revenue per app per month is $821 more through in-app purchasing monetisation compared to advertising-based monetisation (Papas 2013). Even though so many apps monetise through advertising, per user revenue for advertising is less than per user revenue for in-app purchases. The bulk of apps relying upon in-app purchases are predominantly games, and the bulk of app revenue is derived from the gaming sector. Other service-driven apps, such as Uber or Pandora, also feature in-app purchasing.

The split of revenue between the developer and the platform is a key point for app success. For many apps, the platform will command a share of the app revenue. For example, in a game, Google takes 20% of in-app purchase revenue. However, certain apps do not fall victim to this scheme, as their service lies outside the app. Uber, Netflix, Amazon and other firms that offer services through a mobile app avoid paying these fees, as they do not sell app upgrades or digital goods for use within a game.

The Analysis of Mobile App Economics

Most app data is accessible through comprehensive websites built by Apple and Google. The centralised nature of American app stores and similar industrial configurations in other nations enables researchers to gather information about platforms, development firms and products, as such data is publicly available through the app stores and can be accessed through web surfing and scraping, although the amount of data presents a massive data management task. This information can further be supplemented with data available on a subscription basis. Both are discussed below.

Terms and Metrics

Understanding mobile app economics and the success of an app or development firm requires the mastery of a number of relevant terms and analytical concepts. The mobile app industry has its own jargon, and the online world has additional jargon. A number of metrics stand out among the vast quantity of data available from both paid data brokers and scraping from the app marketplaces. To date, no metric offers a perfect understanding of the most successful app. However, when various metrics are used together, the economic impact of an app can be better understood.

Multihoming This generally refers to the ability for one technological entity to relate to multiple other networks, platforms, bases or operating systems. Within the study of mobile app economics, the term refers to the ability of a particular piece of software (app) to run on multiple phone operating systems. For example, as currently constructed, the app Pandora is multihoming because it can run on both Android and iOS in addition to other platforms.

Rank A publicly available rank offers an understanding of overall app success. The specific ranking formulas are proprietary to the platforms. However, industry sources have revealed that the rankings often take into account app download velocity over various periods of time, app ratings and other minor factors. As download velocity can be essentially bought through advertisements or app promotion vendors such as TapJoy, a high rank may not always mean that the app is truly the most successful, used or loved app available. Economics researchers should exercise caution in relying on this measure alone, as many apps achieve a high download velocity while only being used once.

Estimated Revenue Vendors such as App Annie attempt to estimate the revenue gained from in-app purchases by tracking in-app purchase revenue for apps through analytics code that they encourage developers to insert into apps. When app developers do not install the recommended code, App Annie extrapolates revenue based on similar closely ranked apps.

Estimated Total/Average Minutes Spent in App Participant panel-based data vendors such as Nielsen, Arbitron and comScore seek to extrapolate total use time from a group of selected and compensated app testers. This data is of course subject to biases based on sample selection.

Total Downloads On the Android platform, a range of total downloads are displayed per app. Since downloads can be essentially purchased as described above, researchers should always be wary of relying solely on this measure when determining the quality of an app. A reliable source of iOS downloads per app is not publicly available to our knowledge at the time of writing. However, some data firms, such as App Annie, will sell an estimation of extrapolated downloads.

Market Share When focusing on the success of firms or marketplaces, total percentage share of phones sold is useful in crowning a winner. This metric is applicable to both platform firms and handset manufacturers. However, when looking at the economy of one platform, keep abreast of revenue differentials, as the platform with the most users worldwide in 2014 also contains the lowest average revenue per users. Sometimes having fewer users can be more profitable, as is the case for iOS in the USA, as the fewer users have significantly higher average discretionary incomes.

Total Revenue per Platform Industry sources often estimate revenue per platform. The most significant revenue source is in-app purchases. The bulk of this revenue often comes from less than one per cent of game users who spend extraordinary amounts of money improving in-game standing.

Total Unique Users/Visitors This is a count of the number of users that an app has achieved. Researchers should draw a distinction between monthly recurring use and a pure app download. User numbers are important, as often apps are used once before being deleted.

Percentage of Users Reached This metric tracks the total percentage of smartphone owners who use a particular mobile app. It is useful in differentiating the most popular apps from the minor apps. However, it does not take into account revenue or time spent in app.

See Also

- ▶ Internet and the Offline World
- ▶ Online Platforms, Economics of
- ▶ Pricing on the Internet
- ▶ Two-Sided Markets

Bibliography

AppBrain. 2014. *Number of available Android Applications*. http://www.appbrain.com/stats/number-of-android-apps. Accessed 20 Dec 2014.

Bresnahan, T., J.P. Davis, and P. Yin. 2014a. Economic value creation in mobile applications. In *The changing frontier: Rethinking science and innovation*

policy, ed. Adam Jaffe and Benjamin Jones. Chicago: University of Chicago Press.
Bresnahan, T., J. Orsini, and P. Yin. 2014b. Platform non-tipping: Mobile apps. *Working paper*.
Davis, J.P., Y. Muzyrya, and P. Yin. 2014. Experimentation strategies and entrepreneurial innovation: Inherited market differences from the iPhone ecosystem. *Working paper*.
IDC. 2014. *Smartphones to drive double-digit growth of smart connected devices in 2014 and beyond, according to IDC*. IDC Press Release, 17 June. http://www.idc.com/getdoc.jsp?containerId=prUS24935114. Accessed 20 Dec 2014.
O'Toole, J. 2014. *Mobile apps overtake PC internet usage in U.S.* CNN Money. http://money.cnn.com/2014/02/28/technology/mobile/mobile-apps-internet/. Accessed 20 Dec 2014.
Our Mobile Planet. 2013. *United States of America. Understanding the mobile consumer*. Google.
Papas, A. 2013. *Which apps make more money?* http://www.visionmobile.com/blog/ 2013/04/which-apps-make-more-money/. Accessed 20 Dec 2014.
Pocket Gamer. 2014. *App store metrics*. http://www.pocketgamer.biz/metrics/app-store/. Accessed 20 Dec 2014.
US Smartphone Installed Base Share. 2014. http://www.tech-thoughts.net/2012/07/ global-smartphone-market-share-trends.html#.UtdYR_RDtnJ. Accessed 20 Dec 2014.

Mode of Production

R. Jessop

This concept was first introduced by Karl Marx in his efforts to theorize the overall structure and dynamic of capitalism. It has since been widely used, mainly in Marxist political economy and historical studies, to analyse various economic systems. Although there is broad agreement on its general field of application, different approaches exist towards defining and distinguishing particular modes of production. Some of the resulting problems are considered below.

Marx used the concept of mode of production in two main ways: to analyse the economic base and to describe the overall structure of societies. Thus he employed it to specify the particular combination of *forces* and *relations* of production which distinguished one form of labour process and its corresponding form of economic exploitation from another. He also employed it to characterize the overall pattern of social reproduction arising from the relations between the economic base (comprising production, exchange, distribution, and consumption) and the legal, political, social and ideological institutions of the so-called superstructure. The latter usage is particularly problematic. Its conceptual basis is fuzzy and it encourages monocausal economic analyses of whole societies. But even the more rigorously defined and carefully theorized analysis of production proper involves problems. For Marx concentrated on the *capitalist mode of production*, discussed it in relatively abstract terms, and considered precapitalist modes largely in terms of their differences from capitalism. Many of these ambiguities and lacunae survive today so that the meaning and scope of the concept are still contested.

Mode of Production Defined

Marx analysed modes of production in terms of the specific economic form in which the owners of the means of production extracted unpaid surplus labour from the direct producers. For him this form always corresponded to a definite stage of development of the methods of labour and their social productivity. He also described this economic form as 'the innermost secret, the hidden basis of the entire social structure' (*Capital*, III, ch. 47, sect. II). For it provides 'the real foundation on which rise legal and political superstructures and to which correspond definite forms of social consciousness' (1859, Preface). Orthodox Marxists have generally focused on three modes of production: ancient society based on the direct exploitation of slave labour, feudalism with its serf labour and appropriation through ground rent, and capitalism with its free

wage-labour and appropriation through surplus-value (see below).

In general terms a mode of production can be defined as a specific combination of forces and relations of production so organized that it can sustain a distinctive mode of appropriating surplus labour. Forces of production include not only the means and objects of labour but also labour-power itself. They are never purely technical in character but are always shaped by the prevailing social relations of production. The latter can be divided analytically into relations *in* production and relations *of* production (cf. Burawoy 1985). Relations in production comprise the working relations between classes within a productive entity, for example, between capital and labour in the factory; relations of production are grounded in the capacities to allocate resources to diverse productive activities and to appropriate surplus-labour in determinate forms. It is the combination of these forces and relations which defines the basic pattern of class relations and determines the overall pattern of production, distribution, and consumption in its articulation with the appropriation of surplus.

For a distinct mode of production to exist, the forces and relations of production must complement each other so that together they sustain the economic basis of the relevant mode of appropriation. This does not mean that modes of production can somehow reproduce themselves autonomously. There are always extra-economic preconditions (such as law, the state, or specific systems of ideas) which must be secured for economic reproduction to exist. In turn, economic activity is an essential precondition of other activities and its form has its own effects thereon. This mutual presupposition and reciprocal causality have encouraged the extension of the 'mode of production' concept to societies as a whole. Where the forces and relations of production are not mutually supportive and/or their essential extra-economic conditions are not secured, various situations can exist short of an economic collapse. Most studies have examined transitions from one mode of production to another. But is also possible that an *ad hoc*, contingent, and temporary economic system could emerge combining elements from different modes of production.

Some Basic Questions

Almost all the basic questions involved in discussions of modes of production are grounded in the Marxian legacy. Can there be a general theory of modes of production or does each mode have to be examined in its own right? Does a general theory (or even the very concept of mode of production) commit one to an economic reductionist analysis of societies and their succession? How does capitalism differ from (a) pre-capitalist modes and (b) any future communist mode of production? How should one identify the nature and differences among precapitalist modes and, in particular, can one follow Marx in positing a distinctive Asiatic mode of production? Moreover, given that there are different modes of production and forms of labour, how are they to be articulated? How should one periodize the development of particular modes of production? Only some of these issues can be discussed here. Thus no reference will be made to the complex problems involved in defining the modes of production in actually existing or future socialist or communist societies. Likewise only indirect reference will be made to problems of periodization.

A General Theory of Modes of Production?

Orthodox Marxists have followed Marx in dividing economic development into different epochs and in establishing causal links between their economic bases and other social relations. Underpinning this approach there is often a philosophy of history which ascribes an inherent telelogical drive to the sequence of modes of production. This drive is generally attributed to the emergence of a contradiction between the productive forces and the extant relations of production. Whenever the latter hinder the further development of the productive forces, they are overturned through a revolutionary transition to a more progressive

mode (cf. Cohen 1978). In addition to its technological determinism, this approach also suffers from its assumption that only a few pre-capitalist modes existed.

An alternative approach to a general theory was attempted by French structuralists (notably Balibar) in the 1960s and 1970s. Balibar (1970) emphasized the determining role of the relations rather than forces of production and also tried to avoid teleology. He outlined three basic elements and two relations of production to be found in all modes of production and also introduced the concept of a 'transitional mode of production'. Alternative combinations of these constituent elements and relations generated different modes of production. Unfortunately this produced a simplistic and formal taxonomy. It reduced differences among modes to how their constituent elements are combined and thereby implied that the elements themselves are invariant. This ignored the changing social character of both the forces and relations of production. Moreover, whereas the concept of 'transitional' modes is inherently teleological, the idea that all other modes could always reproduce themselves left the problem of historical change unresolved. Thus neither historical materialist orthodoxy nor structuralist taxonomy suggests that a general theory of modes of production is worthwhile.

Are Modes of Production Purely Economic?

This conclusion does not invalidate studies concerned with the structures, genealogies or dynamics of particular modes of production. It means only that these cannot be subsumed under a master theory which explains their specific forms, their succession, and their laws of motion. Particular studies must, of course, define the mode of production under investigation.

There are three main approaches to this task. Firstly, a mode of production can be defined wholly in economic terms, identifying its constituent productive forces and relations of production. Secondly, the forces and relations of production can also be considered in their political and ideological aspects. And, thirdly, the definition can be extended to include the totality of economic, political, and ideological relations necessary for social reproduction as well as economic production. This first definition is unsatisfactory. Pre-capitalist exploitation typically involved extra-economic relations: in turn these could involve direct compulsion (e.g. slavery or the levy of tribute or taxes) and/or political or ideological mechanisms (e.g. a legal monopoly of land or kinship relations). Moreover, not even capitalist production can be reduced to a purely technical process unencumbered by political and ideological considerations. Indeed recent studies have shown the extent to which even the forces of production can embody political and ideological relations by constraining the activities of workers and by maintaining the separation between mental and manual labour. The third definition is also unsatisfactory. For it is equally wrong to include all the political and ideological factors involved in its social reproduction when defining a given mode. This would eliminate the distinction between a mode of production and its extra-economic conditions of existence and thereby encourage neglect of the different ways in which these conditions can be secured. Thus neither a narrow nor a broad definition of modes of production is acceptable.

It is best to consider relations of production as having economic, political, and ideological moments without claiming that they thereby exhaust all social relations. Thus one could study the labour process as involving (a) a socio-technical process of transformation of nature, (b) patterns of coordination, surveillance, and control over workers, and (c) a particular division between mental and manual labour. This does not subsume all political and ideological relations under the mode of production. For, beyond it, there are specific legal, political, social and ideological institutions. How these are articulated with the relations of production (notably through the medium of property relations) will vary from one mode to another. Nor is there any reason to believe that these institutions will always contribute to securing the extra-economic preconditions of a given mode. Finally, nothing in this approach

entails the argument that the forces and/or relations of production determine (whether alone or predominantly) the form and/or content of other social spheres. This has important implications for analysing the unity and coherence of pre-capitalist social formations as well as for the dynamic of class relations more generally.

How Are Modes of Production Articulated?

Granted that different modes of production exist and can be combined, how are they articulated? This question has generated arcane disputes concerning whether two distinct modes of production could co-exist in the same economic space (cf. Wolpe 1982). But is has also led to interesting analyses of the articulation of different forms of social and private labour with a dominant mode of production. These include studies of tribal societies; the impact of capitalism on pre-capitalist societies more generally; the relations between metropolitan and peripheral capital; and the periodization of metropolitan capitalism itself into distinct stages which can be variously combined.

A recent and related topic concerns domestic labour. Some feminists have argued that there is a separate and autonomous domestic mode of production in which women are exploited by a dominant class of men. Others have argued that there is a client domestic mode of production through which capital exploits women because their unpaid domestic labour helps to lower the reproduction costs of all wage-labour. What is clear, such disputes aside, is that domestic labour (as opposed to a domestic mode of production) both contributes to capital accumulation and yet lies beyond it. This highlights the need to examine how modes of production are articulated with other forms of labour.

Pre-Capitalist Modes

Marx and Engels considered pre-capitalist modes of production in several works, most notably in Marx's *Grundrisse* (1857–8). Here Marx suggested an evolutionary schema comprising a tribal stage (with three successive sub-stages, viz. hunting, nomadic pastoralism, and sedentary agriculture); than an ancient slave-holding system based on city-states; then a feudal stage; and then capitalism. He also mentioned Germanic and Slavonic forms of tribalism and outlined an 'Asiatic mode of production'. In all cases he focused on the various forms of agrarian property involved in different modes of production. Marxist economic historians and anthropologists have built on these arguments and have also described other pre-capitalist modes of production.

Ancient Society and Feudalism

Marxists conventionally argue that ancient society was based on slave labour. But slaves can be found in many different economic and political systems so that slavery as such cannot be the defining characteristic of one particular mode of production. It is equally clear that not all the productive labour in ancient societies was performed by slave labour. A better approach emphasizes that ancient societies were organized around city-states and considers how politics intervened in the appropriation of surplus in the ancient mode of production.

Under feudalism a landlord class exploits serf labour. Serfs are tied to the land through political and legal mechanisms and cultivate it on payment of feudal ground rent. Because they actually occupy the land and can determine how it is worked, surplus must be appropriated through customary forms of extra-economic coercion. The particular political shell within which feudal exploitation occurs has often been neglected by Marxist scholars. Yet this makes it difficult to distinguish one form of pre-capitalist rent and its accompanying mode of production from another. It is important to connect more general historical approaches to feudalism (which emphasize such factors as parcellized sovereignty, vassal hierarchy, and the system of economic and military fiefdom) with the analysis of feudal economies.

Only thus can one understand the particular forms and dynamic of feudalism in Europe and Japan as compared with the other agrarian modes of production (cf. Anderson 1974b).

An Asiatic Mode of Production?

Marx provided several different accounts of the Asiatic mode but always emphasized the absence of private property in land. In general he noted that Asiatic societies had autarchic village communities which enjoyed effective communal ownership of the land and which combined crafts with cultivation; but they were also dominated by an overarching state which claimed absolute title to the soil and appropriated the bulk of economic surplus in the form of tax or labour levies.

The scope of this concept seems to vary inversely with that of 'feudalism'. For, given the limited number of modes of production traditionally considered, one or other concept must subsume the most widely divergent economic systems. However, whereas feudalism is generally agreed to be a valid concept and to have been instantiated in the West, neither the concept nor the existence of an Asiatic mode are universally accepted. This partly reflects political disputes concerning the 'semi-Asiatic' character of pre-revolutionary Russia and polemical suggestions that the Soviet system (especially under Stalin) is an Asiatic despotism (e.g. Wittfogel 1957). More generally the concept is theoretically contradictory in Marxist terms (states are not supposed to develop in otherwise classless societies) and also historically inadequate (Asiatic systems were diverse and dynamic rather than homogeneous and stagnant). The history of the concept suggests that there is still much work to be done in analysing pre-capitalist modes of production.

Capitalism

Capitalism involves the generalization of the commodity form to labour-power and the appropriation of surplus-labour in the form of surplus-value. Economic exploitation and capital accumulation both depend upon economic exchange mediated through market forces. This relative separation between economic and extra-economic relations and the dominance of the economic in the dynamic of capital accumulation has encouraged the belief that capitalism can be understood purely as an economic phenomenon. But there are important extra-economic preconditions of capital accumulation (in law, the state, specific forms of family, ideology etc.) and they always intervene in the economic realm. In addition the economic relations themselves have political and ideological moments (cf. above).

Recent studies of relations in production have emphasized how the labour process has important extra-economic aspects. Key concepts here have been the 'politics of production', 'factory regimes', and the mental–manual division of labour (e.g., Burawoy 1985; Thompson 1983). Likewise there have been important non-economistic analyses of capitalist relations of production more generally. Worth noting here are studies concerned with 'regimes of accumulation' and patterns of 'regulation'. These aim to provide a more concrete and conjunctural analysis of capitalist periodization than do more orthodox studies which posit a unilinear and mechanical succession of capitalist stages. They recognize they structural changes and institutional innovation are essential for long-term accumulation and that each national economy has its own specificity within the international system. They emphasize the periodic structural and strategic reorganization of the social relations in and of production. Particular attention has been paid to the shift from regimes based on extensive accumulation to those based on intensive accumulation (especially Fordism). Such studies consider the ensemble of conditions governing the use and reproduction of labour-power, the dynamic of investment and forms of competition, changes in the monetary system, and so on. They also consider changing accumulation strategies and patterns of institutional regulation intended to secure the cohesion

of different national systems and their stable insertion into the international economy (e.g., Aglietta 1979; de Vroey 1984).

Further Research

The concept of mode of production is clearly both complex and problematic. This is particularly true for pre-capitalist modes. Studies here have frequently adhered too rigidly to Marx's own typologies and also find difficulty in handling the intimate connections between their precapitalist relations of production and extra-economic relations. But there is enormous scope for further research on pre-capitalist modes. In dealing with capitalist economies, the most promising areas of research comprise: (a) the politics of production and associated 'factory regimes'; (b) regimes of accumulation and patterns of regulation; and (c) the articulation of capitalism with other modes of production and/or forms of social or domestic labour. In each case this means paying more careful and systematic attention to the articulation between the economic, political, and ideological moments of production. Without progress in this direction the spectres of teleology, technological determinism, and monocausal economic explanations will continue to haunt Marxist analyses.

See Also

- ▶ Agricultural Growth and Population Change
- ▶ Capital as a Social Relation
- ▶ Capitalism
- ▶ Economic Interpretation of History
- ▶ Industrial Revolution
- ▶ Peasants
- ▶ Socially Necessary Technique

Bibliography

Aglietta, M. 1979. *A theory of capitalist regulation*. London: New Left Books.

Anderson, P. 1974a. *Passages from antiquity to feudalism*. London/New York: New Left Books/Schocken.

Anderson, P. 1974b. *Lineages of the absolutist state*. London/New York: New Left Books/Schocken.

Balibar, E. 1970. The basic concepts of historical materialism. In *Reading capital*, ed. L. Althusser and E. Balibar. London/New York: New Left Books/Pantheon.

Banaji, J. 1977. Modes of production in a materialist conception of history. *Capital and Class* 3: 1–44.

Bloch, M. 1961. *Feudal society.* 2 vols. London/Chicago: Routledge & Kegan Paul/University of Chicago Press.

Bloch, M. 1982. *Marxism and anthropology.* London: Oxford University Press.

Burawoy, M. 1985. *The politics of production*. London: New Left Books.

Cohen, G.A. 1978. *Karl Marx's theory of history*. Oxford: Oxford University Press.

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

Finley, M. 1982. *Economy and society in ancient Greece*. London: Chatto & Windus.

Harvey, D. 1982. *The limits of capital*. Oxford/Chicago: Blackwell/Chicago University Press.

Hindess, B., and P.Q. Hirst. 1975. *Pre-capitalist modes of production*. London: Routledge & Kegan Paul.

Hobsbawm, E. 1964. Introduction. In *Pre-capitalist economic formations*, ed. K. Marx. London: Lawrence & Wishart.

Kula, W. 1976. *Economic theory of the feudal system*. London: New Left Books.

Marx, K. 1859. Preface to *A contribution to the critique of political economy.* London: Lawrence & Wishart, 1971.

Marx, K. 1867–94. *Capital.* 3 vols. Harmondsworth: Penguin.

Marx, K. 1957–8. *Grundrisse der Kritik der politischen Oekonomie.* Berlin: Dietz. Trans., Harmondsworth: Penguin, 1973.

Molyneux, M. 1979. Beyond the domestic labour debate. *New Left Review* 116:3–27.

Padgug, R.A. 1976. Problems in the theory of slavery and slave society. *Science and Society* 40(1): 3–27.

de St. Croix, G. 1982. *The class struggle in the ancient Greek world*. London/Ithaca: Duckworth/Cornell University Press.

Thompson, P. 1983. *The nature of work*. London: Macmillan.

Turner, B.S. 1978. *Marx and the end of Orientalism*. London: Macmillan.

de Vroey, M. 1984. A regulation approach interpretation of the contemporary crisis. *Capital and Class* 23: 45–66.

Wittfogel, K.A. 1957. *Oriental despotism: A comparative study of total power.* New Haven: Yale University Press.

Wolpe, H. (ed.). 1982. *The articulation of modes of production*. London: Routledge & Kegan Paul.

Model Averaging

Gernot Doppelhofer

Keywords

Bayes' rule; Bayesian estimation; Bayesian model averaging; Empirical Bayes methods; Exchangeability; Frequentist model averaging; Homoskedasticity; Likelihood; Markov chain Monte Carlo methods; Metropolis–Hastings algorithm; Model averaging; Model selection criteria; Model uncertainty; Posterior model probabilities; Sensitivity analysis; Statistical decision theory; Stochastic search variable selection

JEL Classifications

C10; C50; D81; E52; O40

Model averaging allows the estimation of the distribution of unknown parameters and related quantities of interest across different models. The basic principle of model averaging is to treat models and associated parameters as unobservable and estimate their distributions based on observable data. Model averaging can be employed for inference, prediction and policy analysis in the face of model uncertainty. Many areas of economics give rise to model uncertainty, including uncertainty about theory, specification and data issues. A naive approach that ignores model uncertainty generally results in biased parameter estimates, overconfident (too narrow) standard errors and misleading inference and predictions (see Draper 1995). Taking model uncertainty seriously implies a departure from conditioning on a particular model and calculating quantities of interest by averaging across different models instead.

Model averaging is conceptually straightforward. The sample information contained in the likelihood function for a particular model is combined with relative model weights or posterior model probabilities to estimate the distribution of unknown parameters across models. Three main approaches – Bayesian, empirical Bayes, and frequentist – have been developed, and they differ in their underlying statistical foundations and practical implementation.

Bayesian model averaging (*BMA*) was developed first to systematically deal with model uncertainty. The idea of combining evidence from different models is readily integrated into a Bayesian framework. Jeffreys (1961) laid the foundation for BMA, further developed by Leamer (1978). Hoeting et al. (1999), Wasserman (2000) and Koop (2003) give excellent introductions to BMA. A drawback of the Bayesian approach is that it requires assumptions about prior information about distribution of unknown parameters. In response, *empirical Bayes* (*EB*) approaches have been developed to estimate elements of the prior using observable data. Chipman et al. (2001) argue for a pragmatic approach that introduces objective or frequentist considerations into model averaging. In contrast to Bayesian approaches, *frequentist model averaging* (*FMA*) methods were developed only relatively recently. Recent contributions include Yang (2001), Hjort and Claeskens (2003) and Hansen (2007).

Model averaging was not widely used until advances in statistical techniques and computing power facilitated its practical use (see Chib 2001; Geweke and Whiteman 2006). Economic applications of model averaging include economic growth (Fernandez et al. 2001a; Sala-i-Martin et al. 2004), finance (Avramov 2002), policy evaluation (Brock et al. 2003; Levin and Williams 2003), macroeconomic forecasting (Garratt et al. 2003).

This article is organized as follows. The statistical model averaging framework is introduced in the next section. Different model averaging approaches are illustrated with applications to linear regressions. Finally, implementation issues, including model priors, numerical methods, and software are discussed.

Statistical Framework

Suppose a decision maker observes data Y and wishes to learn about quantities of interest related to an unknown parameter (vector) θ, such as the effect of an economic variable (say $\theta > 0$ or $\theta \leq 0$) or predictions of future observations Y^f. The utility (or loss) function of the decision maker describes the relation between parameter of interest θ and action a. For example, the decision maker could maximize expected utility

$$\max_a E[u(a,\theta|Y)] = \int u(a,\theta|Y)p(\theta|Y)d\theta. \quad (1)$$

In general, the preferred action depends on the preferences of the decision-maker and the unconditional distribution of parameters. Alternative preference structure can have important consequences for optimal estimators and implied policy conclusions. Bernardo and Smith (1994) give an accessible introduction to statistical decision theory. In the context of economic policy, Brock et al. (2003) present an interesting discussion of alternative preferences and implied policies.

A key ingredient in decision making is the *posterior* distribution of the parameter θ, which can be calculated using Bayes's rule:

$$p(\theta|Y) = \frac{L(Y|\theta)p(\theta)}{p(Y)} \propto L(Y|\theta)p(\theta). \quad (2)$$

The posterior distribution is therefore proportional to the *likelihood* function $L(Y|\theta)$, which summarizes all information about θ contained in the observed data, and the *prior* distribution $p(\theta)$. In contrast, the classical approach assumes that the parameter θ is fixed (non-random) and does not have a meaningful distribution. The estimator $\hat{\theta}$ on the other hand is viewed as a random variable.

In many economic and more generally non-experimental applications, a decision maker might face considerable model uncertainty given potentially overlapping, economic theories. Brock and Durlauf (2001) refer to this as 'open-endedness' of economic theories. Also, there might be alternative empirical specifications of these theoretical channels. In sum, the number of observations may be smaller than the number of suggested explanations, and the problem may be compounded by data problems, such as missing data or outliers.

Formally, there may be many candidate models $M_1,\ldots,M_K$ to explain the observed data. A model M_j can be described by a probability distribution $p(Y|\theta_j, M_j)$ with model-specific parameter (vector) θ_j. In a situation of model uncertainty, the decision-maker evaluates the utility function Eq. (1) using the posterior distribution of θ. The posterior distribution is unconditional with respect to the set of models and is calculated by averaging conditional or model-specific distributions across all models

$$p(\theta|Y) = \sum_{j=1}^{K} w_j \cdot p(\theta_j|M_j,Y), \quad (3)$$

where the model weights w_j are proportional to the fit in explaining the observable data. In a Bayesian context, the weights are the *posterior model probabilities,* $w_j = p(M_j|Y)$. Using Bayes's rule,

$$p(M_j|Y) = \frac{L(Y|M_j)p(M_j)}{\sum_{j=1}^{K} L(Y|M_j)p(M_j)} \propto L(Y|M_j)p(M_j). \quad (4)$$

The posterior model weights are proportional to the product of prior model probability $p(M_j)$ and model-specific marginal likelihood $L(Y|M_j)$. The marginal likelihood is obtained by integrating a model-specific version of equation Eq. (2) with respect to θ_j

$$L(Y|M_j) = \int_{\theta} L(Y|\theta_j,M_j)p(\theta_j|M_j)d\theta_j \quad (5)$$

using the fact that $\int p(\theta_j|M_j, Y)d\theta_j = 1$.

When comparing two models, M_i and M_j say, the posterior model probabilities or *posterior odds ratio* equals the ratio of integrated likelihoods times the prior odds

$$\frac{p(M_i|Y)}{p(M_J|Y)} = \frac{L(Y|M_i)p(M_i)}{L(Y|M_j)p(M_j)}. \quad (6)$$

Similarly, the weight for model M_i relative to K models under consideration is given by Eq. (4), where the normalizing factor $\sum_{j=1}^{K} L(Y|M_j)\, p(M_j)$ ensures consistency of model weights.

The decision maker may be interested in particular aspects of the unconditional distribution Eq. (3), such as posterior mean or variance. Leamer (1978) derives the following expressions for unconditional mean or variance of the parameter θ

$$E(\theta|Y) = \sum_{j=1}^{K} p(M_j|Y) E(\theta_j|Y, M_j). \quad (7)$$

$$\begin{aligned} Var(\theta|Y) &= E(\theta^2|Y) - [E(\theta|Y)]^2 \\ &= \sum_{j=1}^{K} p(M_j|Y) \\ &\times \{Var(\theta_j|Y, M_j) + [E(\theta_j|Y, M_j)]^2\} \\ &- [E(\theta|Y)]^2 \quad (8) \\ &= \sum_{j=1}^{K} p(M_j|Y) Var(\theta_j|Y, M_j) \\ &+ \sum_{j=1}^{k} p(M_j|Y) [E(\theta_j|Y, M_j) - E(\theta|Y)]^2. \end{aligned}$$

The expression for the unconditional mean of θ in Eq. (7) is simply the model-weighted sum of conditional means. Notice that the unconditional variance of θ in Eq. (8) exceeds the sum of model-weighted conditional variances by an additional term, reflecting the distance between the estimated conditional mean in each model $E(\theta_j|Y, M_j)$ and the unconditional mean $E(\theta|Y)$. Ignoring this last term overestimates the precision of estimated effects and underestimates parameter uncertainty (see Draper 1995).

The advantage of the Bayesian approach to model averaging is its generality and the explicit treatment of model uncertainty and decision theory. The decision maker simply combines prior information about the distribution of parameters and models with sample information to calculate the unconditional posterior distribution of θ in Eq. (3).

However, there are several problems that can make implementation of BMA difficult in practice (see Hoeting et al. 1999; Chipman et al. 2001):

1. The specification of prior distribution of parameters θ requires assumptions about functional forms and unknown hyper-parameters which will in general affect the marginal likelihood Eq. (5) and hence posterior model weights Eq. (4).
2. The specification of prior probabilities over the model space $p(M_j)$ might have important effects on posterior model weights Eq. (4).
3. The number of models K in Eq. (3) can be too large for a complete summation across models, implying the use simulation techniques to approximate the unconditional distribution $p(\theta|Y)$ in equation Eq. (3).
4. Choices of utility function Eq. (1) and class of models are other important issues.

These issues are discussed in turn, contrasting the fully Bayesian, empirical Bayes and frequentist approaches.

Linear Regression Example

Many of the implementation problems of model averaging and approaches suggested in the literature can be illustrated using the linear regression example (see Koop 2003). Raftery et al. (1997) and Fernandez et al. (2001b) discuss BMA for linear regression models.

Consider linear regression models of the form

$$y = x_1\beta_1 + \cdots + x_k\beta_k + \varepsilon = X\beta + \varepsilon, \quad (9)$$

where y is the vector of N observations of the dependent variable and $X = [x_1, \ldots, x_k]$ is a set of k regressors (including a constant) with associated coefficient vector β. Each model M_j is characterized by a subset of explanatory variables X_j with coefficient vector β_j. With k regressors, the set of linear models equals $K = 2^k$. The residuals are drawn from a multivariate normal distribution and are assumed to be conditionally homoskedastic, $\varepsilon_j \sim N(0, \sigma^2 I)$. Notice that this implies that the residuals are also conditionally exchangeable (see Bernardo and Smith 1994; Brock and Durlauf 2001).

Suppose the decision maker is interested in the effect of different explanatory variables, represented by slope parameters β with posterior distribution of $p(\beta|Y)$. As shown in Eq. (3), the posterior distribution is estimated by weighting conditional distributions of parameters by posterior model probabilities. The relative posterior model weights in Eqs (6) and (4) are proportional to the marginal likelihood and prior model weights.

For the normal regression model, the likelihood function can be written as

$$\begin{aligned} &L(y|\beta_j,\sigma^2) \\ &=\frac{1}{(2\pi\sigma^2)^{N/2}}\left\{\exp\left[-\frac{1}{2\sigma^2}(y-X\beta_j)'(y-X\beta_j)\right]\right\} \\ &\times\propto\left\{\exp\left[-\frac{1}{2\sigma^2}\left(\beta_j-\hat{\beta}_j\right)'X_j'X_j\left(\beta_j-\hat{\beta}_j\right)\right]\right\} \\ &\times\left\{\sigma^{-(v_j+1)}\exp\left[-\frac{v_j s_j^2}{2\sigma^2}\right]\right\}. \end{aligned} \tag{10}$$

The second line of the likelihood substitutes the ordinary least squares (OLS) estimates for the slope and variance

$$\hat{\beta}_j=\left(X_j'X_j\right)^{-1}X_j'y, \tag{11}$$

$$s_j^2=\frac{\left(y-X_j\hat{\beta}_j\right)'\left(y_j-X_j\hat{\beta}_j\right)}{v_j}, \tag{12}$$

with degrees of freedom $v_j=N-k_j-1$. The implementation of model averaging – Bayesian, empirical Bayes, or frequentist – requires the specification of prior distributions $p(\theta_j)$ for the model parameters $\theta_j=(\beta_j,\sigma^2)$.

Bayesian Conjugate Priors

A standard way to specify priors in Bayesian estimation is to assume a prior structure that is analytically and computationally convenient. A *conjugate prior* distribution leads to a posterior distribution of the same class of distributions when combined with the likelihood. The likelihood Eq. (10) is part of the Normal-Gamma family of distributions, proportional to the product of a normal distribution for the slope β_j, conditional on the variance σ^2, and an inverse-Gamma distribution for the variance σ^2. The conjugate prior therefore takes the form

$$\begin{aligned} &p(\beta_j|\sigma^2,M_j)\sim N(\beta_{0j},\sigma^2V_{0j})p(\sigma^2|M_j) \\ &=p(\sigma^2)\sim IG(s_0^2,v_0) \end{aligned} \tag{13}$$

where the prior hyper-parameters for slope and variance are denoted by subscript 0. Notice that the error variance is assumed to be drawn from the same distribution across all regression models, reflecting the assumption of conditional homoskedasticity and exchangeability of the residuals.

A drawback of the Bayesian approach is that marginal likelihood and posterior model weights depend on unknown hyper-parameters (β_0, V_0, s_0, v_0). Different subjective priors therefore affect the posterior model weights and distribution of parameters, and hence also the decision maker's action. The standard Bayesian approach to check for robustness with respect to the choice of prior parameters is sensitivity analysis. An alternative strategy is to limit the use of subjective prior information and use objective methods based on observed data.

Empirical Bayes Priors

Empirical Bayes (EB) approaches make use of sample information to specify prior parameters. Different versions of empirical Bayes methods have been proposed in the literature (see Hoeting et al. 1999; George and Foster 2000; Chipman et al. 2001). To limit the importance of prior information, EB methods often use non-informative or diffuse priors that are dominated by the sample information (see Leamer 1978). Jeffreys (1961) proposes non-informative priors to represent lack of prior knowledge and derives a formal relationship to the expected information in the sample.

A drawback of non-informative priors is that they are usually not proper distributions, which

can lead to undesirable properties when comparing models with different parameters. In this case, relative model weights can depend on arbitrary constants. However, this problem is not present when comparing models with common parameters, since normalizing constants drop out from *relative* model weights (see Kass and Raftery 1995). Koop (2003) argues that informative or proper priors should be used for all other (non-common) parameters.

Fernandez et al. (2001b) propose *benchmark priors* for BMA that limit the subjective prior information to a minimum while maintaining the Bayesian natural conjugate framework. They suggest the following non-informative priors for the error variance, assumed to be the same in all k models:

$$p(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{14}$$

The slope parameter β_j is drawn from a normal prior distribution as in Eq. (13) with prior mean $\beta_{0j} = 0$ and prior covariance matrix V_{oj} equal to the so-called g-prior suggested by Zellner (1986):

$$V_{oj} = \left(g_0 X_j' X_j\right)^{-1}. \tag{15}$$

Intuitively, the prior covariance matrix is assumed to be proportional to the sample covariance with a factor of proportionality g_0. The g-prior simplifies the specification of prior covariances to choosing a single parameter g_0. For example, $g_0 = 0$ corresponds to completely non-informative priors, and $g_0 = 1$ implies a very informative prior receiving equal weight to the sample information. Based on extensive simulations, Fernandez et al. (2001b) recommend the following benchmark values:

$$g_0 = \begin{cases} 1/k^2, & \text{if} \quad N \le k^2 \\ 1/N, & \text{if} \quad N > k^2 \end{cases}. \tag{16}$$

Note that the ratio of prior to sample variance g_0 decreases with the sample size or with the square of estimated parameters. If the number of parameters is relatively large $k^2 \ge N$, the variance is assumed to be relatively more diffuse.

Using this prior structure, the posterior weights for model M_j can be written as

$$p(M_j|Y) \propto p(M_j) \cdot \left(\frac{1+g_0}{g_0}\right)^{-k_j/2} \cdot SSE_j^{-(N-1)/2}. \tag{17}$$

The weight for model $p(M_j|Y)$ depends on three terms: (i) the prior model weight $p(M_j)$, (ii) a penalty term for the number of regressors $((1+g_0)/g_0)^{-k_j/2}$ implying a preference for parsimonious models, and (iii) a term involving the sum of squared errors of the regression $SSE_j \equiv (y - X_j\beta_j)'(y - X_j\beta_j)$, corresponding to the kernel of the normal likelihood.

Frequentist Sample Dominated Priors

A potential problem of using non-informative g-priors for the error covariance matrix is that the limit of posterior weights may be very sensitive to specification of the prior (see Leamer 1978). Alternatively, Leamer (1978) assumes that a proper, conjugate Normal-Gamma prior Eq. (13) is 'dominated' by the sample information as the number of observations N grows. For stationary regressors with $\lim_{N\to\infty}(X'_j\ X_j)/N$ converging to a constant, the implied model weight is approximately equal to the (exponentiated) Schwarz (1978) model selection criterion (BIC)

$$p(M_j|Y) \propto p(M_j) \cdot N^{-k_j/2} \cdot SSE_j^{-N/2}. \tag{18}$$

On closer inspection, the relative model weights using non-informative g-priors Eq. (17) or sample-dominated prior Eq. (18) are essentially the same, using $g_0 = 1/N$ in Eq. (16). This is very reassuring for a decision maker, since the relative model weights are very similar under an empirical Bayesian or frequentist interpretation.

The BIC weights can also be derived from a unit information prior, where the information introduced by the prior corresponds to *one* datapoint from the sample (see Kass and Wasserman 1995; Raftery 1995). Klein and Brown (1984) give an

alternative derivation of the BIC model weights Eq. (18) by minimizing the so-called

Shannon information in the prior distribution; this approach also lends support for using the BIC model weights in small samples.

The underlying model space and its interpretation are important issues in the model uncertainty literature. Bernardo and Smith (1994) distinguish between *M*closed and *M*-open environments, where the former includes the true model and the latter does not necessarily. A set of Akaike (AIC) model weights can be derived in the *M*-open environment as the best approximation to the true distribution (see Burnham and Anderson 2002). The AIC weights have the disadvantage that they will not be consistent in *M*-closed environments.

Prior Over Model Space

An important ingredient to model averaging is the choice of prior model probability. A popular choice is to impose a uniform prior over the space of models

$$p(M_j) = 1/K. \tag{19}$$

This prior might represent diffuse information about the set of models, but does have important implications for the size of models.

There are different approaches to modelling the inclusion of explanatory variables in the linear regression models Eq. (9). Mitchell and Beauchamp (1988) assign a discrete prior probability mass $p(\beta_i = 0|M_j)$ to excluding regressors x_i from the regression model M_j, that is a 'spike' at zero. A more Bayesian approach assigns a mixture of a relatively informative prior at zero (corresponding to a spike at zero) and a more diffuse prior if the variable is included (see George and McCulloch 1993).

An alternative to specifying prior model probabilities is to think about prior model size and the implied probability of including individual variables. Sala-i-Martin et al. (2004) argue that in the context of economic growth regressions a prior model size $\overline{k}$ smaller than the one implied by uniform priors $k/2$ might be preferable. Notice that this translates into a prior probability $\pi = p\left(\beta_i \neq 0 | M_j\right) = \overline{k}/k$ of including a regressor x_i in model M_j. The implied model probability can then be written as

$$p(M_j) = \pi^{k_j} \cdot (1-\pi)^{k-k_j}. \tag{20}$$

Notice that the prior inclusion probabilities π_i and implied prior model weights can also differ across variables, which is used in the 'stratified' sampler of the BACE approach by Sala-i-Martin et al. (2004) to speed up numerical convergence.

George (1999) observes that, when allowing for a large number of explanatory variables which could be correlated with each other, posterior model probabilities can be spread across models with 'similar' regressors. To address this problem, George (1999) proposes *dilution priors*, which reduce the prior weight on models that include explanatory variables measuring similar underlying theories. Alternatively, one can impose a hierarchical structure on the set of models and variables and partition the model space accordingly (see Chipman et al. 2001; Brock et al. 2003). Doppelhofer and Weeks (2007) propose to estimate the degree of dependence or jointness among regressors over the model space. If we are only interested in prediction, the orthogonalization of regressors greatly reduces the computational burden of model averaging (see Clyde et al. 1996). The costs are the loss of interpretation of associated coefficient estimates and the need to recalculate orthogonal factors with changing sample information.

Numerical Simulation Techniques

A major challenge for the practical implementation of model averaging is the computational burden of calculating posterior quantities of interest when the model space is potentially very large. In the linear regressions example, an exhaustive integration over all 2^k models becomes impractical for a relatively moderate number of 30 regressors.

Recent advances in computing power and development of statistical methods have made

numerical approximations of posterior distributions feasible. Chib (2001) gives an overview of computationally intensive methods. Such methods include Markov chain Monte Carlo techniques (Madigan and York 1995), stochastic search variable selection (George and McCulloch 1993), the Metropolis–Hastings algorithm (Chib and Greenberg 1995), and the Gibbs sampler (Casella and George 1992). Chipman et al. (2001) contrast different approaches in the context of Bayesian model selection.

The main idea of Monte Carlo simulation techniques is to estimate the empirical distribution of the parameter θ or related functions of interest $g(\theta)$ by sampling from the posterior distribution

$$E[g(\theta)|Y] = \int g(\theta)p(\theta|Y)d\theta, \qquad (21)$$

where $g(\theta)$ could be any function, such as variance of θ or predicted values of the dependent variable y. Consider the sample counterpart

$$\hat{g}_S = \frac{1}{S}\sum_{s=1}^{s} g\left(\theta^{(s)}\right), \qquad (22)$$

where $\theta^{(s)}$ is a random *i.i.d.* sample drawn from $p(\theta|Y)$ and S is the number of draws. Provided that $E[g(\theta)|Y] < 0$ exists, a weak law of large numbers implies

$$\hat{g}_S \xrightarrow{p} E[g(\theta)|Y]. \qquad (23)$$

A central limit theorem implies that

$$\sqrt{S}\{\hat{g}_S - E[g(\theta)|Y]\} \xrightarrow{d} N\left(0, \Sigma_g\right) \qquad (24)$$

where Σ_g is the estimated covariance matrix of $g(\theta)|Y$.

Markov chain Monte Carlo (MCMC) techniques strengthen these results by constructing a Markov chain moving through the model space $\{M(s), s = 1,\ldots, S\}$ that simulates from a transition kernel $p(\theta^{(s)}|\theta^{(s-1)})$, starting from an initial value $\theta^{(0)}$. There are various approaches to constructing a Markov chain that converges to the posterior distribution $p(\theta|Y)$. This limiting distribution can be estimated from simulated values of $\theta^{(s)}$.

Simulation methods differ with respect to the choice of sampling procedure and transition kernels. A sampling algorithm that uses the underlying structure of the model can greatly improve the efficiency of the simulation. For example, the Gibbs sampler uses the structure of the statistical model to partition parameters and their distribution into blocks, which breaks up the simulation into smaller steps. In the linear regression example, the Gibbs sampler can draw from the conditional distributions for slope and variance parameters Eq. (13) separately. A disadvantage of numerical methods can be the technical challenges in their implementation (for an excellent introduction, see Gilks et al. 1996). Links to software packages and codes that facilitate implementation, such as BACC, BACE, BUGS and the BMA project website, are listed at the end of this article.

An alternative approach is to limit the set of models and rule out dominated models by Occam's razor, see Hoeting et al. (1999). This can speed up computation of posterior distributions and can be useful tool for model selection. Evidence by Raftery et al. (1996) suggests that model averaging leads to important improvements in predictive performance over any single model, and gives a small predictive advantage relative to the restricted set of models. The relative performance of different model averaging techniques and associated model weights depends on sample size and stability of estimated model (see Yuan and Yang 2005; Hansen 2007).

See Also

- Bayesian Econometrics
- Bayesian Methods in Macroeconometrics
- Bayesian Statistics
- Decision Theory in Econometrics
- Econometrics
- Extreme Bounds Analysis
- Hierarchical Bayes Models
- Model Uncertainty
- Shrinkage-Biased Estimation in Econometrics
- Testing

Bibliography

Avramov, D. 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics* 64: 423–458.

Bernardo, J.M., and A.F.M. Smith. 1994. *Bayesian theory.* New York: Wiley.

Brock, W.A., and S.N. Durlauf. 2001. Growth empirics and reality. *World Bank Economic Review* 15: 229–272.

Brock, W.A., S.N. Durlauf, and K. West. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1): 235–322.

Burnham, K.P., and D.R. Anderson. 2002. *Model selection and multimodel inference: A practical information-theoretic approach*. 2nd ed. New York: Springer.

Carlin, B.P., and T.A. Louis. 2000. *Bayes and empirical Bayes Methods for data analysis*. 2nd ed. New York: Chapman & Hall.

Casella, G., and E.I. George. 1992. Explaining the Gibbs sampler. *The American Statistician* 46: 167–174.

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

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

Chipman, H., E.I. George, and R.E. McCulloch. 2001. The practical implementation of Bayesian model selection. In *Model selection. IMS lecture notes: Monograph series*, ed. P. Lahiri. Beachwood: Institute of Mathematical Statistics.

Clyde, M., H. Desimone, and G. Parmigiani. 1996. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* 91: 1197–1208.

Doppelhofer, G. and M. Weeks. 2007. Jointness of growth determinants. *Journal of Applied Econometrics*.

Draper, D. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* 57: 45–97.

Fernandez, C., E. Ley, and M.F.J. Steel. 2001a. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576.

Fernandez, C., E. Ley, and M.F.J. Steel. 2001b. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100: 381–427.

Garratt, A., K. Lee, M.H. Pesaran, and Y. Shin. 2003. Forecast uncertainties in macroeconomic modelling: An application to the U.K. economy. *Journal of the American Statistical Association* 98: 829–838.

George, E.I. 1999. Discussion of Bayesian model averaging and model search strategies by M.A. Clyde. *Bayesian Statistics* 6: 175–177.

George, E.I., and D.P. Foster. 2000. Calibration and empirical Bayes variable selection. *Biometrika* 87: 731–747.

George, E., and R.E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88: 881–889.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57: 1317–1339.

Geweke, J., and C. Whiteman. 2006. Bayesian forecasting. In *Handbook of economic forecasting*, ed. G. Elliott, C.W.J. Granger, and A. Timmermann, vol. 1. Amsterdam: North-Holland.

Gilks, W., S. Richardson, and D. Spiegelhalter. 1996. *Markov Chain Monte Carlo in practice*. New York: Chapman & Hall.

Hansen, B.E. 2007. Least squares model averaging. *Econometrica* 75: 1175–1189.

Hjort, N.L., and G. Claeskens. 2003. Frequentist model averaging. *Journal of the American Statistical Association* 98: 879–899.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14: 382–417.

Jeffreys, H. 1961. *Theory of probability.* 3rd ed. Oxford: Clarendon Press.

Kass, R.E., and A.E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795.

Kass, R.E., and L. Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90: 928–934.

Klein, R.W., and S.J. Brown. 1984. Model selection when there is 'minimal' prior information. *Econometrica* 52: 1291–1312.

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

Leamer, E. 1978. *Specification searches*. New York: Wiley.

Levin, A.T., and J.C. Williams. 2003. Robust monetary policy with competing reference models. *Journal of Monetary Economics* 50: 945–975.

Madigan, D., and J. York. 1995. Bayesian graphical models for discrete data. *International Statistical Review* 63: 215–232.

Mitchell, T.J., and J.J. Beauchamp. 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83: 1023–1032.

Raftery, A.E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25: 111–163.

Raftery, A.E., D. Madigan, and J.A. Hoeting. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179–191.

Raftery, A.E., D. Madigan, and C.T. Volinsky. 1996. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* 5: 323–349.

Sala-i-Martin, X., G. Doppelhofer, and R.M. Miller. 2004. Determinants of economic growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94: 813–835.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Wasserman, L. 2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology* 44: 92–107.

Yang, Y. 2001. Adaptive regression by mixing. *Journal of the American Statistical Association* 96: 574–588.
Yuan, Z., and Y. Yang. 2005. Combining linear regression models: When and how? *Journal of the American Statistical Association* 100: 1202–1214.
Zellner, A. 1986. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti*, ed. P.K. Goel and A. Zellner. Amsterdam: North-Holland.

Model Averaging Software and Codes

BACC package: http://www2.cirano.qc.ca/Bbacc
BACE website: http://www.nhh.no/sam/bace
BMA homepage: http://www.research.att.com/Bvolinsky/bma.html
BUGS project: http://www.mrc-bsu.cam.ac.uk/bugs
LeSage's Econometrics Toolbox: http://www.spatial-econometrics.com

Model Selection

Jean-Marie Dufour

Abstract

The problem of statistical model selection in econometrics and statistics is reviewed. Model selection is interpreted as a decision problem through which a statistical model is selected in order to perform statistical analysis, such as estimation, testing, confidence set construction, forecasting, simulation, policy analysis, and so on. Broad approaches to model selection are described: (1) hypothesis testing procedures, including specification and diagnostic tests; (2) penalized goodness-of-fit methods, such as information criteria; (3) Bayesian approaches; (4) forecast evaluation methods. The effect of model selection on subsequent statistical inference is also discussed.

Keywords

ARMA models; Autocorrelation; Bayesian statistics; Deterministic models; Econometrics; Endogeneity; Forecasting; Forecast evaluation; Heteroskedasticity; Linear models; Model selection; Models; Parsimony; Probability models; Serial correlation; Specification problems in econometrics; Statistical decision theory; Statistical inference; Stochastic models; Structural change; Testing; Time series analysis

JEL Classifications

C10; C50; D81; E52; O40

The purpose of econometric analysis is to develop mathematical representations of observable phenomena, which we call *models* or *hypotheses* (models subject to restrictions). Such models are then used to perform parameter estimation, test hypotheses, build confidence sets, make forecasts, conduct simulations, analyse policies, and so on. A central feature of modelling activity is the fact that models are usually interpreted as stylized (or simplified) representations that can perform certain tasks – such as prediction – but (eventually) not others, and they are treated *as if they were true* for certain purposes. Indeed, summarizing and stylizing observed phenomena can be viewed as essential components of modelling activity, which make it useful. This feature is not specific to economics and is shared by other sciences (see Cartwright 1983).

Models can be classified as either *deterministic* or *stochastic*. *Deterministic models*, which often claim to make arbitrarily precise predictions, can be useful in theoretical activity. However, such models are rarely viewed as appropriate representations of observed data; for example, unless they are highly complex or indeterminate, they are typically logically inconsistent with data. For this reason, models used for econometric analysis are usually *stochastic* (or *statistical*).

Formally, a statistical model is a family of probability distributions (or measures) which are proposed to represent observed data. *Model selection*, in this context, is the task of selecting a family of proposed probability distributions, which will then be used to analyse data and perform other statistical inference operations (such as parameter estimation, hypothesis testing, and so on).

A basic feature of probability models is that they are typically *unverifiable*: as for any theory

that makes an indefinite number of predictions, we can never be sure that the model will not be at odds with new data. Moreover, they are *logically unfalsifiable*: in contrast with deterministic models, a probabilistic model is usually logically compatible with all the possible observation sets. Consequently, model selection can depend on a wide array of elements, such as the objectives of the model, (economic) theory, the data themselves, and various conventions.

Features which are often viewed as desirable include: (*a*) simplicity or *parsimony* (Zellner et al. 2001); (*b*) the ability to deduce testable (or falsifiable) hypotheses (Popper 1968); (*c*) the possibility of interpreting model parameters in terms of economic theory, if not consistency with economic theory; (*d*) the ability to satisfactorily perform the tasks for which the model is built (prediction, for example); (*e*) consistency with observed data. It is important to note that these characteristics depend (at least, partially) on conventional elements, such as the objectives of the model, criteria upon which a model will be deemed 'satisfactory', and so on. For further discussions of these general issues, the reader may consult Poirier (1994), Morgan and Morrison (1999), Keuzenkamp (2000), Zellner et al. (2001) and Dufour (2003).

In this article, we focus on statistical methods for selecting a model on the basis of the available data. Methods for that purpose can be classified in four broad (not mutually exclusive) categories:

1. Hypothesis testing procedures, including specification and diagnostic tests;
2. Penalized goodness-of-fit methods, such as information criteria;
3. Bayesian approaches;
4. Forecast evaluation methods.

The three first approaches are meant to be applicable 'in-sample', while the last approach *stricto sensu* requires observations that are not available when the model is selected, but may lead to model revision. (For general reviews of the topic of statistical model selection in econometrics and statistics, see Hocking 1976; Leamer 1978, 1983; Draper and Smith 1981; Judge et al. 1985, chs 7 and 21; Sakamoto et al. 1985; Grasa 1989; Choi 1992; Gouriéroux and Monfort 1995, ch. 22; Charemza and Deadman 1997; McQuarrie and Tsai 1998; Burnham and Anderson 2002; Clements and Hendry 2002; Miller 2002; Bhatti et al. 2006). It is also interesting to note that classification techniques in statistics contain results that may be relevant to model selection. This topic, however, goes beyond the scope of the present article (for further discussion, see Krishnaiah and Kanal 1982).

Model Selection and Specification Errors

Most model selection methods deal in different ways with a trade-off between model *realism* – which usually suggests considering relatively general, hence complex models – and *parsimony*. From the viewpoint of estimation, for example, a model which is too simple (or parsimonious) involves *specification errors* and *biases* in parameter estimation, while too complex a model leads to parameter estimates with large variances. If the objective is forecasting, it is usually unclear which effect dominates.

For example, let us consider a linear regression model of the form

$$y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tk}\beta_k + u_t, \quad t = 1, \ldots, T, \tag{1}$$

where y_t is a dependent variable and x_{t1},.. .,x_{tk} are explanatory variables, and u_t is a random disturbance which is typically assumed to be independent of (or uncorrelated with) the explanatory variables. In the classical linear model, it is assumed that the regressors can taken as fixed and that the disturbances u_1,.. .,u_T are independent and identically distributed (i.i.d.) according to a $N(0,\sigma^2)$ distribution. In this context, model selection typically involves selecting the regressors to be included as well as various distributional assumptions to be made upon the disturbances.

An especially important version of (1) is the autoregressive model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t, \quad t = 1, \ldots, T. \tag{2}$$

Then a central model selection issue consists in setting the order p of the process. In such models, there is typically little theoretical guidance on the order, so data-based order selection rules can be quite useful. A related set-up where model selection is usually based on statistical methods is the class of autoregressive-moving-average (ARMA) models

$$y_t = \beta_0 + \phi y_{t-1} + \cdots + \phi p y_{t-p} + u_t - \theta_1 u_{t-1} + \cdots + \theta_1 u_{t-q}, \tag{3}$$

where the orders p and q must be specified.

By considering the simple linear regression model, it is easy to see that excluding irrelevant variables can lead to biases in parameter estimates (Theil 1957). On the other hand, including irrelevant regressors raises the variances of the estimators. The overall effect on the mean square error (MSE) of the estimator and, more generally, how closely it will tend to approach the parameter value may be ambiguous. It is well known that a biased estimator may have lower MSE than an unbiased estimator. This may be particularly important in forecasting, where a simple 'false' model may easily provide better forecasts than a complicated 'true' model, because the latter may be affected by imprecise parameter estimates (Allen 1971).

Hypothesis Testing Approaches

Since hypothesis tests are based on a wide body of statistical theory (see Lehmann 1986; Gouriéroux and Monfort 1995), such procedures are widely used for assessing, comparing and selecting models. Furthermore, econometric models are also based on economic theory which suggests basic elements that can be used for specifying models. This entails a form of asymmetry, in which restrictions suggested by economic theory will be abandoned only if 'sufficient evidence' becomes available. Although significance tests are meant to decide whether a given hypothesis (which usually takes the form of a restricted model) is compatible with the data, such procedures can also be used for model selection. It is interesting to note that the methodology originally proposed by Box and Jenkins (1976) for specifying ARMA models was almost exclusively based on significance tests (essentially, autocorrelation tests).

There are two basic ways of using hypothesis tests for that purpose. The first one is *forward* or *specific-to-general* approach, in which one starts from a relatively simple model and then checks whether the model can be deemed 'satisfactory'. This typically involves various specification tests, such as:

1. Residual-based tests, including tests for heteroskedasticity, autocorrelation, outliers, distributional assumptions (for example, normality), and so on;
2. Tests for unit roots and/or stationarity, to decide whether corrections for integrated variables may be needed;
3. Tests for the presence of structural change;
4. Exogeneity tests, to decide whether corrections for endogeneity – such as instrumental variable (IV) methods – are required;
5. Tests for the addition of explanatory variables;
6. Tests of the functional form used (for example, linearity vs. nonlinearity).

There is a considerable literature on specification tests in econometrics (see Godfrey 1988; MacKinnon 1992; Davidson and MacKinnon 1993). Systematic procedures for adding variables are also know in statistics as *forward selection* or *stepwise regression* procedures (Draper and Smith 1981).

The second way is the *backward* or *general-to-specific* approach, in which one starts from a relatively comprehensive model which includes all the relevant variables. This model is then simplified by checking which variables are significant. *Backward selection* procedures in statistics (Draper and Smith 1981) and the general-to-specific approach in econometrics (Davidson et al. 1978; Charemza and Deadman 1997) can be viewed as illustrations of this approach.

In practical work, the backward and forward approaches are typically combined. Both involve a search for a model which is both parsimonious and consistent with the data. However, the results may differ. Specifying a model through significance tests involves many judgements and depends on idiosyncratic decisions. Further, standard hypothesis tests involve the use of typically conventional levels (such as the commonly used five per cent level). The powers of the tests can also have a strong influence on the results.

Penalized Goodness-of-Fit Criteria

As pointed out by Akaike (1974), it is not clear that hypothesis testing is a good basis for model selection. Instead, the problem of model selection may be better interpreted as an estimation problem involving a well-defined loss function. This leads to the topic of goodness-of-fit criteria.

A common way of assessing the performance of a regression model, such as (1), consists in computing the coefficient of determination, that is, the proportion of the dependent variable variance which is 'explained' by the model:

$$R^2 = 1 - \frac{\hat{V}(u)}{\hat{V}(y)} \tag{4}$$

where $\hat{V}(u) = \sum_{t=1}^{T} \hat{u}_t^2/T, \hat{V}(y) = \sum_{t=1}^{T} (y_t - \overline{y})^2/T, \overline{y} = \sum_{t=1}^{T} y_t y_t/T$ and $\hat{u}_1, \ldots, \hat{u}_T$ least squares residuals. This measure, however, has the inconvenient feature that it always increases when a variable is added to the model, even if it is completely irrelevant, and it can be made equal to its maximal value of one by including a sufficient number of regressors (for example, using any set of T linearly independent regressors).

An early way of avoiding this problem was proposed by Theil (1961, p. 213) who suggested that $\hat{V}(u)$ and $\hat{V}(y)$ be replaced by the corresponding unbiased estimators $s^2 = \sum_{t=1}^{T} \hat{u}_t^2/(T-k)$ and $s_k^2 = \sum_{t=1}^{T} (y_t - \overline{y})^2/(T-1)$. This yields the adjusted coefficient of determination:

$$\begin{aligned}\overline{R}^2 &= -\frac{s^2}{s_y^2} = 1 - \frac{T-1}{T-k}(1-R^2) \\ &= R^2 - \frac{k-1}{T-k}(1-R^2).\end{aligned}$$

It is easy to see that $\overline{R}^2$ may increase when the number of regressors increases. Note that maximizing $\overline{R}^2$ is equivalent to minimizing the 'unbiased estimator' s^2 of the disturbance variance. Further, if two regression models (which satisfy the assumptions of the classical linear model) are compared, and if one of these is the 'true' model, then the value of s^2 associated with the true model is smaller on average than the one of the other model (see Theil 1961, p. 543). On the other hand, in large samples, the rule which consists in maximizing $\overline{R}^2$ does not select the true model with a probability converging to one: that is, it is not consistent (see Gouriéroux and Monfort 1995, section 2.3).

Another approach consists in evaluating the 'distance' between the selected model and the true (unknown) model. Let $f(y)$ the density associated with the postulated model and $f_o(y)$ the density of the true model, where $Y = (y_1, \ldots, y_T)'$. One such distance is the *Kullback distance*:

$$\begin{aligned}I(f, f_0) &= \int \log[f_0(y)/f(y)] f_0(y)\,dy \\ &= \underset{f_o}{E}\{\log[f_o(Y)/f(Y)]\} \\ &= \underset{f_o}{E}\{\log[f_o(Y)]\} - \underset{f_o}{E}\{\log[f(Y)]\}.\end{aligned}$$

Minimizing $I(f, f_o)$ with respect to f is equivalent to minimizing $-\underset{f_o}{E}\{\log[f(Y)]\}$. We obtain an information criterion by selecting an 'estimator' of $\underset{f_o}{E}\{\log[f(Y)]\}$.

For the case where the model is estimated by maximizing a likelihood function $L_T(\theta)$ over a $K \times 1$ parameter vector θ, Akaike (1973) suggests that $L\left(\hat{\theta}\right)$ can be viewed as a natural estimator of $\underset{f_o}{E}\{\log[f(Y)]\}$. However, the fact that θ has been estimated introduces a bias. This bias is (partially) corrected – using an expansion argument – by

subtracting the number K from $L\left(\hat{\theta}\right)$. This suggests the following information criterion:

$$AIC_L\left(\hat{\theta}_T\right) = -2L_T\left(\hat{\theta}_T\right) + 2K \qquad (5)$$

where K is the dimension of θ (the number of estimated parameters) and multiplication by 2 is introduced to simplify the algebra. Among a given set of models, the one with the lowest AIC is selected.

The above criterion has also been generalized by various authors leading the following general class of criteria:

$$IC_L\left(\hat{\theta}_T\right) = -2L_T\left(\hat{\theta}_T\right) + c(T,K)K \qquad (6)$$

where $c(T, K)$ is a function of T and K. In the case of Gaussian models, such as (1) or (2) with i.i.d. $N(0,\sigma^2)$ disturbances, we have $L_T\left(\hat{\theta}_T\right) = -(T/2)\ln\left(\hat{\sigma}_T^2\right) + d_T$, where d_T is a constant which only depends on T, so that minimizing $IC_L\left(\hat{\theta}_T\right)$ is equivalent to minimizing

$$IC\left(\hat{\theta}_T\right) = \ln\left(\hat{\sigma}_T^2\right) + c(T,K)\frac{K}{T}. \qquad (7)$$

Alternative values of $c(T, K)$ which have been proposed include:

1. $c(T, K) = 2$ (Akaike 1969), which yields what is usually called the AIC criterion;
2. $c(T, K) = \ln(T)$ (Schwarz 1978);
3. $c(T, K) = 2\delta_T \ln(\ln T)$ where $\limsup_{T\to\infty} \delta_T > 1$ (Hannan and Quinn 1979);
4. $c(T,K) = 2 + \frac{2K(K+1)}{T-K-1}$ (Hurvitch and Tsai 1989), which leads to the AIC criterion.

An especially convenient feature of such information criteria is the fact that they can be applied to both regression models (through (7)) as well as to various nonlinear models (using (6)).

Other related rules include: (*a*) criteria based on an estimate of the *final prediction error*, which try to estimate the mean square prediction error taking into account estimation uncertainty (Akaike 1969, 1970; Mallows 1973; Amemiya 1980); (*b*) the *criterion autoregressive transfer* (CAT) function proposed by Parzen (1977) for selecting the order of an autoregressive process; (*c*) Sawa's (1978) Bayesian information criterion (BIC).

By far, the information criteria are the most widely used in practice. Some theoretical (non-) optimality properties have been established. In particular, when one of the models compared is the 'true' one, it was observed by Shibata (1976) that Akaike's criterion is not consistent, in the sense that it does not select the most parsimonious true model with probability converging to one (as the sample size goes to infinity). Instead, even in large samples it has a high probability of picking a model with 'too many parameters'. By contrast, the criterion proposed by Hannan and Quinn (1979) is consistent under fairly general conditions, which also entails that Schwarz's (1978) criterion also leads to consistent model selection. On the other hand, the AIC criterion has a different optimality property, in the sense that it tends to minimize the one-step expected quadratic forecast error (Shibata 1980).

On consistency, it is also interesting to observe that consistent model selection rules can be obtained provided each model is tested through a consistent test procedure (against all the other models considered) and the level of the test declines with the sample size at an appropriate rate (which depends on the asymptotic behaviour of the test statistic) (see Pötscher 1983).

Model selection criteria of the information have the advantage of being fairly mechanical. On the other hand, they can be become quite costly to apply in practice when the number of models considered is large.

Bayesian Model Selection

Bayesian model selection involves comparing models through their 'posterior probabilities' giving observed data. Suppose we have two models M_1 and M_2 each of which postulates that the observation vector y follows a probability density which depends on a parameter vector: $p_y(y|\theta_1, M_1)$ under M_1, and $p_y(y|\theta_2, M_2)$ under M_2, where θ_1 and

θ_2 are unknown parameter vectors (which may have different dimensions). Further, each one of the parameter vectors is assigned a 'prior distribution' ($p(\theta_1|M_1)$ and $p(\theta_2|M_2)$), and each model a 'prior probability' ($p(M_1)$ and $p(M_2)$). Then one may compute the 'posterior probability' of each model given the data

$$p(M_i|\,y) = p(M_i)\int p_y(y|\,\theta_i, M_1)p(\theta_1|\,M_i)\mathrm{d}\theta\, i, \quad i = 1, 2. \tag{8}$$

This posterior probability of each model provides a direct measure of the 'plausibility' of each model. In such contexts, the ratio

$$K_{12} = \frac{p(M_1|\,y)}{p(M_2|\,y)} \tag{9}$$

is called the 'posterior odds ratio' of M_1 relative to M_2.

A rational decision rule for selecting between M_1 and M_2 then emerges if we can specify a loss function such as

$$L(i,j) = \text{costofchoosing}\, M_j\, \text{when}\, M_i\, \text{istrue}. \tag{10}$$

If $L(i, i) = 0$ for $i = 1, 2$, expected loss is minimized by choosing M_1 when

$$K_{12} \geq \frac{L(2,1)}{L(1,2)}, \tag{11}$$

and M_2 when otherwise. In particular, if $L(1, 2) = L(2, 1)$, expected loss is minimized by choosing the model with the highest posterior probability. Such rules can be extended to problems where more than two models are compared.

The Bayesian approach automatically introduces a penalty for non-parsimony and easily allows the use of decision-theoretic considerations. The main difficulty consists in assigning prior distributions on model parameters and prior probabilities to competing models. For further discussion, see Zellner (1971, ch. 10), Leamer (1978, 1983), Gelman et al. (2003) and Lancaster (2004).

Forecast Evaluation

In view of the fact that forecasting is one of the most common objectives for building econometric models, alternative models are often assessed by studying *post-sample* forecasts. Three types of assessments are typically considered in such contexts: (*a*) tests of predictive failure; (*b*) descriptive measures of forecast performance, which can be compared across models; (*c*) tests of predictive ability.

A test of predictive failure involves testing whether the prediction errors associated with a model are consistent with the model. This suggests testing whether forecasts are 'unbiased' or 'too large' to be consistent with the model. The well-known predictive test for structural change proposed by Chow (1960) is an early example of such an approach. (For further discussion and extensions, see Box and Tiao 1976; Dufour 1980; Pesaran et al. 1985; Dufour et al. 1994; Dufour and Ghysels 1996; Clements and Hendry 1998.)

Common measures of forecast performance involve mean errors, mean square errors, mean absolute errors, and so on (see Theil 1961; Diebold 2004). Although commonly used, such measures are mainly descriptive. They can usefully be complemented by tests of predictive ability. Such procedures test whether the difference between expected measures of forecast performance is zero (or less than zero) against an alternative where it is different from zero (or larger than zero). Tests of this type were proposed, among others, by Meese and Rogoff (1988), Diebold and Mariano (1995), Harvey et al. (1997), West (1996), West and McCracken (1998) and White (2000) (for reviews, see also Mariano 2002; McCracken and West 2002).

It is important to note that predictive performance and predictive accuracy depend on two features: first, whether the theoretical model used is close to the unknown data distribution and, second, the ability to estimate accurately model parameters (hence on sample size available for estimating these). For a given sample size, a false but parsimonious model may well have better predictive ability than the 'true' model.

Post-model Selection Inference

An important issue often raised in relation with model selection is its effect on the validity of inference – such as estimation, tests and confidence sets – obtained after a process of model selection (or pretesting). This issue is subtle and complex. Not surprisingly, both positive and negative assessments can be found.

On the positive side, it has been observed that pretesting (or model selection) does allow one to produce so-called 'super-efficient' (or Hodges) estimators, whose asymptotic variance can be at least as low as the Cramér–Rao efficiency bound and lower at certain points (see Le Cam 1953). This may be viewed as a motivation for using consistent pretesting.

Furthermore, consistent model selection does not affect the asymptotic distributions of various estimators and test statistics, so the asymptotic validity of inferences based on a model selected according to such a rule is maintained (see Pötscher 1991; Dufour et al. 1994).

On the negative side, it is important to note that these are only asymptotic results. In particular, these are pointwise convergence results, not uniform convergence results, so they may be quite misleading concerning what happens in finite samples (for some examples, see Dufour 1997; Pötscher 2002). For estimation, there is a considerable literature on the finite-sample distribution of pretest estimators, which can be quite different of their limit distributions (Judge and Bock 1978; Danilov and Magnus 2004). For a critical discussion of the effect of model selection on tests and confidence sets, see Leeb and Pötscher (2005).

Conclusion

The problem of model selection is one of the most basic and challenging problems of statistical analysis in econometrics. Much progress has been done in recent years in developing better model selection procedures and for understanding the consequences of model selection.

But model building remains largely an art in which subjective judgements play a central role. Developing procedures applicable to complex models, which may involve a large number of candidate variables, and allowing for valid statistical inference in the presence of model selection remain difficult issues to which much further research should be devoted.

See Also

- ▶ Bayesian Statistics
- ▶ Econometrics
- ▶ Endogeneity and Exogeneity
- ▶ Forecasting
- ▶ Heteroskedasticity and Autocorrelation Corrections
- ▶ Linear Models
- ▶ Models
- ▶ Serial Correlation and Serial Dependence
- ▶ Specification Problems in Econometrics
- ▶ Statistical Decision Theory
- ▶ Statistical Inference
- ▶ Structural Change
- ▶ Testing
- ▶ Time Series Analysis

M

Bibliography

Akaike, H. 1969. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21: 243–247.

Akaike, H. 1970. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22: 203–217.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, ed. B.N. Petrov and F. Csaki. Budapest: Akademiai Kiado.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716–723.

Allen, D.M. 1971. Mean square error prediction as a criterion for selecting variables. *Technometrics* 13: 469–475.

Amemiya, T. 1980. Selection of regressors. *International Economic Review* 21: 331–354.

Bhatti, M.I., H. Al-Shanfari, and M.Z. Hossain. 2006. *Econometric analysis of model selection and model testing*. Aldershot: Ashgate.

Box, G.E.P., and G.M. Jenkins. 1976. *Time series analysis: Forecasting and control*. 2nd ed. San Francisco: Holden-Day.

Box, G.E.P., and G.C. Tiao. 1976. Comparison of forecast and actuality. *Applied Statistics* 64: 195–200.

Burnham, K.P., and D.R. Anderson. 2002. *Model selection and multi-model inference: A practical information theoretic approach*. New York: Springer.

Cartwright, N. 1983. *How the laws of physics lie*. Oxford: Oxford University Press.

Charemza, W.W., and D.F. Deadman. 1997. *New directions in econometric practice: General to specific modelling, cointegration and vector autoregression*. 2nd ed. Aldershot: Edward Elgar.

Choi, B. 1992. *ARMA model identification*. New York: Springer.

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

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

Clements, M.P., and D.F. Hendry, ed. 2002. *A companion to economic forecasting*. Oxford: Blackwell.

Danilov, D., and J.R. Magnus. 2004. On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122: 27–46.

Davidson, J.E.H., D.F. Hendry, F. Srba, and S. Yeo. 1978. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88: 661–692.

Davidson, R., and J.G. MacKinnon. 1993. *Estimation and inference in econometrics*. New York: Oxford University Press.

Diebold, F.X. 2004. *Elements of forecasting*. 3rd ed. Mason, OH: Thomson South-Western.

Diebold, F.X., and R.S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13: 253–263.

Draper, N.R., and H. Smith, ed. 1981. *Applied regression analysis*. rev ed. New York: Wiley.

Dufour, J.-M. 1980. Dummy variables and predictive tests for structural change. *Economics Letters* 6: 241–247.

Dufour, J.-M. 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. *Econometrica* 65: 1365–1389.

Dufour, J.-M. 2003. Identification, weak instruments and statistical inference in econometrics. *Canadian Journal of Economics* 36: 767–808.

Dufour, J.-M. and Ghysels, E. 1996. Recent developments in the econometrics of structural change. *Journal of Econometrics* 70(1).

Dufour, J.-M., E. Ghysels, and A. Hall. 1994. Generalized predictive tests and structural change analysis in econometrics. *International Economic Review* 35: 199–229.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2003. *Bayesian data analysis*. 2nd ed. London: Chapman and Hall/CRC.

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

Gouriéroux, C., and A. Monfort. 1995. *Statistics and econometric models*. Vol. 1 and 2. Cambridge: Cambridge University Press.

Grasa, A.A. 1989. *Econometric model selection: A new approach*. Dordrecht: Kluwer.

Hannan, E.J., and B. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41: 190–191.

Harvey, D.I., S.J. Leybourne, and P. Newbold. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13: 281–291.

Hocking, R.R. 1976. The analysis and selection of variables in linear regression. *Biometrika* 32: 1–49.

Hurvitch, C.M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76: 297–307.

Judge, G.G., and M.E. Bock. 1978. *The statistical implications of pre-test and stein-rule estimators in econometrics*. Amsterdam: North-Holland.

Judge, G.G., W.E. Griffiths, R. Carter Hill, H. Lütkepohl, and T.-C. Lee. 1985. *The theory and practice of econometrics*. 2nd ed. New York: Wiley.

Keuzenkamp, H.A. 2000. *Probability, econometrics and truth: The methodology of econometrics*. Cambridge: Cambridge University Press.

Krishnaiah, P.R., and L.N. Kanal. 1982. *Handbook of statistics 2: Classification, pattern recognition and reduction of dimensionality.* Amsterdam: North-Holland.

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

Le Cam, L. 1953. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics* 1: 277–330.

Leamer, E. 1978. *Specification searches: ad hoc inferences with nonexperimental data*. New York: Wiley.

Leamer, E.E. 1983. Model choice. In *Handbook of econometrics*, ed. Z. Griliches and M.D. Intriligator, Vol. 1. Amsterdam: North-Holland.

Leeb, H., and B. Pötscher. 2005. Model selection and inference: Facts and fiction. *Econometric Theory* 21: 29–59.

Lehmann, E.L. 1986. *Testing statistical hypotheses*. 2nd ed. New York: Wiley.

MacKinnon, J.G. 1992. Model specification tests and artificial regressions. *Journal of Economic Literature* 30: 102–146.

Mallows, C.L. 1973. Some comments on *Cp*. *Technometrics* 15: 661–675.

Mariano, R.S. 2002. Testing forecast accuracy. In *Clements and Hendry* (2002).

McCracken, M.W. and West, K.D. 2002. Inference about predictive ability. In *Clements and Hendry* (2002).

McQuarrie, A.D.R., and C.-L. Tsai. 1998. *Regression and time series model selection*. Singapore: World Scientific.

Meese, R.A., and K. Rogoff. 1988. Was it real? The exchange rate-interest differential relation over the modern floating-rate period. *Journal of Finance* 43: 933–948.

Miller, A. 2002. *Subset selection in regression*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.
Morgan, M.S., and M. Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.
Parzen, E. 1977. Multiple time series: Determining the order of approximating autoregressive schemes. In *Multivariate analysis IV: Proceedings of the fourth international symposium on multivariate analysis*, ed. P.R. Krishnaiah. Amsterdam: North-Holland/ Elsevier.
Pesaran, M.H., R.P. Smith, and J.S. Yeo. 1985. Testing for structural stability and predictive failure: a review. *Manchester School* 3: 280–295.
Poirier, D.J. 1994. *The methodology of econometrics*. Vol. 2. Aldershot: Edward Elgar.
Popper, K. 1968. *The logic of scientific discovery*. rev ed. New York: Harper Torchbooks.
Pötscher, B.M. 1983. Order estimation in ARMA-models by Lagrange multiplier tests. *Annals of Statistics* 11: 872–885.
Pötscher, B. 1991. Effects of model selection on inference. *Econometric Theory* 7: 163–185.
Pötscher, B. 2002. Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots and estimation of long memory parameters. *Econometrica* 70: 1035–1065.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1985. *Akaike information criterion statistics*. Dordrecht: Reidel.
Sawa, T. 1978. Information criteria for discriminating among alternative regression models. *Econometrica* 46: 1273–1291.
Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 71: 117–126.
Shibata, R. 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8: 147–164.
Theil, H. 1957. Specification errors and the estimation of economic relationships. *Review of the International Statistical Institute* 25: 41–51.
Theil, H. 1961. *Economic forecasts and policy*. 2nd ed. Amsterdam: North-Holland.
West, K.D. 1996. Asymptotic inference about predictive ability. *Econometrica* 64: 1067–1084.
West, K.D., and M.D. McCracken. 1998. Regression-based tests of predictive ability. *International Economic Review* 39: 817–840.
White, H. 2000. A reality check for data snooping. *Econometrica* 68: 1097–1126.
Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.
Zellner, A., H.A. Keuzenkamp, and M. McAleer. 2001. *Simplicity, inference and modelling: Keeping it sophisticatedly simple*. Cambridge: Cambridge University Press.

Model Uncertainty

Alexei Onatski

Abstract

Model uncertainty is a condition of analysis when the specification of the model of analysed process is open to doubt. A failure to account for model uncertainty may result in poor decisions. This article reviews various approaches to representing model uncertainty. The approaches depend on the research context, differ in their degree of generality, and may be classified as deterministic versus stochastic, Bayesian versus frequentist, and treating model uncertainty as static versus viewing model uncertainty as evolving over time.

Keywords

Ambiguity aversion; Approximating model; Axioms of choice; Bayesian statistics; Bootstrap; Detection error probability; Dynamic stochastic general equilibrium models; Ellsberg paradox; ε-contaminated priors; Heterogeneity uncertainty; Identification scheme; Identified vector autoregressions; Incommensurability; Incomplete information; Judgement; Laplace transform; Learning; Linear quadratic Gaussian problem; Markov processes; Model averaging; Model expansion; Model uncertainty; Model uncertainty set; Models; Monetary policy; Overconfidence; Probability; Rational expectations; Relative entropy; Risk-sensitive control; Specification uncertainty; Theory uncertainty; Uncertainty; Vector autoregressions

JEL Classifications

C10; C50; D81; E52; O40

Model uncertainty is a condition of analysis when the specification of the model of analysed process is open to doubt. One of the fundamental sources

of model uncertainty is the tradition of critical reasoning that, in words of Karl Popper (1962, pp. 151–153), 'admits a plurality of doctrines which all try to approach the truth by means of critical discussion'. Popper traces the critical tradition back to ancient Greek philosophy. He cites Xenophanes, 570–480 BC, who wrote (see Diels 1951, vol. 1, pp. 133 and 137):

> The gods did not reveal, from the beginning,
> All things to us; but in the course of time,
> Through seeking, men find that which is the better. . .
>
> But as for certain truth, no man has known it,
> Nor will he know it; neither of the gods,
> Nor yet of all the things of which I speak.
>
> And even if by chance he were to utter
> The final truth, he would himself not know it:
> For all is but a woven web of guesses.

Another fundamental source of model uncertainty is the necessity for models to be simple enough to provide an efficient link between theory and reality (see Morgan and Morrison 1999, for a book-length discussion of the nature of models). Complicated models may be less useful than simple ones even though the accuracy of the simple models' description of the modelled process may be more doubtful.

The understanding of what constitutes a model and how to model model uncertainty itself depends on the research context. For example, for engineers a prototypical model is represented by a system of differential equations:

$$\dot{x}(t) = Ax(t) + Bu(t),$$
$$y(t) = Cx(t) + Du(t),$$

where $x(t)$, $u(t)$, and $y(t)$ are square-integrable functions of $t \in [t\,,\ \infty\ )$, interpreted, respectively, as internal states, inputs and outputs of the modelled mechanism. By means of the Laplace transform, we get an alternative representation of the above model (with zero initial states):

$$\hat{y}(s) = M(s)\hat{u}(s),\ \ M(s) = D + C(sI - A)^{-1}B,$$

where $\hat{u}(s)$ and $\hat{y}(s)$ are the Laplace transforms of $u(t)$ and $y(t)$ (see Kwakernaak and Sivan 1972, p. 33).

For engineers, the interest often lies in checking whether particular inputs into the modelled mechanism are such that the corresponding internal states and outputs satisfy an admissibility criterion. Since the model is only an approximation of the mechanism, the check of the admissibility criterion should take into account possible deviations of the model from the truth. These possible deviations constitute model uncertainty, which can be represented by a set of models built around the above reference model.

A model of the uncertainty set which is very flexible and well suited for the purpose of the admissibility checks is the so-called linear fractional model (see Zhou et al. 1996, Chaps. 10 and 11). It replaces the reference model represented by $M(s)$, by a set of models:

$$M(\Delta) = M(s) + L(s)\Delta(I - \Delta G(s))^{-1}R(s), \Delta \in \Lambda,$$

where Λ is a set of block-diagonal matrices with the largest singular value bounded by unity. The number and the structure of the blocks, and the form of matrix functions $L(s)$, $R(s)$, and $G(s)$ must be chosen so that the resulting model uncertainty set accurately represents the engineer's understanding of possible deviations of the reference model from the modelled mechanism.

To take another example, for researchers in statistics a model is often defined as a family of the joint probability distributions of the data. Draper (1995, pp. 45–46) notes that statistical models can be expressed in two parts, the first representing structural assumptions, such as distributional choices for residuals, or a particular functional form of the regression function and so on, and the second representing parameters, whose meaning is specific to the assumed structure. He points out that 'even in controlled experiments and randomized sample surveys key aspects of . . . [the structure] will usually be uncertain, and this is even more true with observational studies'. Statistical model uncertainty can be interpreted as the structural uncertainty that Draper is concerned about in the above citation.

A failure to account for statistical model uncertainty often leads to overconfidence in the results of a statistical study. For example, the forecast intervals, which are computed ignoring possible model uncertainty, may be too narrow, p-values of a test of significance of coefficients in a linear regression too small, and so forth.

Typically, statistical models analyse reality, which is much more complicated than man-made mechanisms. Therefore, building a crisp set of models that represent model uncertainty is more problematic in statistics than in engineering. Most of the statistical approaches to modelling model uncertainty are Bayesian. They represent model uncertainty by a prior distribution defined in the model space and propagate this uncertainty to the statistical decisions by integrating models out from the posterior distribution. Such a technique of model uncertainty propagation is called Bayesian model averaging (see Hoeting et al. 1999 for a tutorial).

There are no standardized ways of specifying a prior that would represent model uncertainty. One approach is to expand a given model to a more general class and to formulate a subjective prior over this class. An early example of this approach can be found in Box and Tiao (1962), who re-examine Darwin's paired data on the heights of self- and cross-fertilized plants earlier analysed by Fisher (1935). To take into account a possible misspecification of Fisher's model for differences in the heights of the i-th pair of plants, y_i:

$$y_i = \theta + \sigma e_i, \quad e_i \propto i.i.d.N(0,1),$$

Box and Tiao expand it to a more general model:

$$y_i = \theta + \sigma e_i, \quad e_i \text{ are i.i.d}$$
$$\text{with density proportional to } \exp\left(-\frac{1}{2}|e|^{2/(1+\beta)}\right),$$

and formulate a beta-type prior distribution for the 'extra' parameter β.

As emphasized by Draper et al. (1987, p. 12), for the model expansion approach to be successful it is important to 'stake out the corners in model space', that is, to find 'the plausible variations on the model... that strongly influence what actions would be taken'. Of course, such an exercise would necessarily be subjective and context specific.

An interesting frequentist alternative to specifying a prior in the model space is to bootstrap the modelling process. Efron and Gong (1983) consider a databased process of explanatory variable selection for a logistic model of the probability of death from a certain disease. They apply the selection process to bootstrap replications of the data, obtaining, thus, a distribution of logistic models, which represents uncertainty about the model whose explanatory variables were chosen on the basis of the original data-set.

If we turn to a discussion of model uncertainty in economics, we first note that the economic model uncertainty is much broader in scope than engineering or statistical model uncertainty. The economic reality is so complex that it may be impossible in principle to approximate it by any model. Different research communities may disagree on what should be understood by economic reality in the first place. For example, Frankel and Rockett (1988, p. 318), in their study of potential gains from cooperation of different countries' monetary authorities, write: 'The assumption that policy makers agree on the true model has little, if any, empirical basis. Different governments subscribe to different economic philosophies.'

The idea of incommensurability of different views of economic reality is a focus of Dow's (2004) methodological study of model uncertainty. Dow puts the incommensurability idea in the context of uncertainty research originating in Keynes's (1921) *Treatise on Probability*, and discusses the role of judgement in a situation when it is impossible in principle to compare models on the basis of their closeness to 'the truth'.

Further, in views of Keynes (1921) and Knight (1921), economic uncertainty may be conceptually different from the uncertainty modelled by randomness. So even if the incommensurability issue does not arise, an economic decision-maker may be hesitant in assigning probabilities to different economic models and comparing them on the basis of these probabilities. Such a view is

supported by a range of experimental studies initiated by the Ellsberg paradox.

The Ellsberg paradox shows that people prefer to bet on 50–50 lotteries rather than on lotteries with completely unknown odds. Such behaviour is a variant of a more general phenomenon called ambiguity aversion. It reveals that people fail to assign prior probabilities to events that happen in incomplete information environments.

As Gilboa and Schmeidler (1989) show, failing to assign prior probabilities to events is perfectly rational because it is consistent with axioms of choice as reasonable as those used by Savage (1954). Gilboa and Schmeidler's axioms imply that a rational decision-maker acts as if he or she contemplates a set of probability distributions over the possible events. The decision is then made so as to minimize the expected loss under the worst possible distribution from the set.

Much of modern research on economic model uncertainty concerns monetary policy formulation and evaluation when policymakers do not have a single reliable model of the economy. In the rest of this article we will therefore focus on the model uncertainty arising in monetary policy research.

Global Model Uncertainty

Different approaches to macroeconomic model uncertainty can roughly be separated in two broad categories, which Brock et al. (2003) call global and local approaches. The global approach assumes that a set of possible models consists of the substantially different economic theories. An early example of the global approach is posited by McCallum (1988), who uses a real business cycle model, a monetary misperception model and a Keynesian model to represent model uncertainty confronted by a monetary policymaker. In contrast, the local approach builds the model uncertainty set by continuously expanding a single reference model.

Brock et al. (2003) distinguish three different components of model uncertainty in the global approach. The first component is 'theory uncertainty', which represents economists' 'disagreement over fundamental aspects of the economy'. The second component is 'specification uncertainty', which includes uncertainty about lag length specification, functional form, and the choice of proxy variables representing particular theoretical concepts. The last component is 'heterogeneity uncertainty' which 'concerns the extent to which different observations are assumed to obey a common model'.

A model for the global model uncertainty itself is based on a set of models which represent different theories, have different specifications given a particular theory, and may include dummy variables or other devices that capture possible data heterogeneity. Brock et al. (2003) propose to complete the model of model uncertainty by specifying a prior distribution over the models in the set. They propose three principles that should guide the formulation of the prior. First, it 'should assign relatively high probability to those areas of the likelihood that are relatively large' (see, however, Chris Sims's critique of this principle in the discussion published with Brock et al. 2003). Second, 'a prior should be robust in the sense that a small change in the prior should not induce a large change in the posterior'. Finally, 'priors should be flexible enough to allow for their use across similar studies'.

To accommodate the possibility of the ambiguity aversion on behalf of policymakers, Brock et al. (2003) suggest that the chosen prior, π, be extended to a class of ε-contaminated priors $\{(1-\varepsilon)\pi + \varepsilon P, P \in P(M)\}$, where $0 \leq \varepsilon \leq 1$ and $P(M)$ is the set of all possible probability measures on the model uncertainty set. The policy which takes into account the model uncertainty can then be chosen by minimizing the expected posterior loss under the worst possible prior from the ε-contaminated class.

Classes of ε-contaminated priors are often used in robust Bayesian analysis to model uncertainty in the prior distribution (see Berger and Berliner 1986). Such classes are easy to work with, and it is not difficult to show that the policy described above differs from the policy which minimizes expected posterior loss under the original prior by putting an extra weight on the worst possible model from the model uncertainty set. The higher the ε, the larger the extra weight.

In the extreme case when $\varepsilon = 1$, specifying prior probabilities of the models from the model uncertainty set is not necessary. The very set completely describes model uncertainty. The policy is then formulated as if the worst possible model were true. The policy choice under the extreme model uncertainty is often visualized as a zero-sum game between a policymaker and malevolent 'nature' who chooses adversary models from the model uncertainty set. Early advocates of this useful visualization were Brunner and Meltzer (1969) and von zur Muehlen (2001).

Describing model uncertainty by an unweighted set of models was advocated by John Tukey. He says in his comment on Draper (1995) (see Draper 1995, p. 78):

> The most acceptable pattern, as far as I am concerned, for the development of a bouquet of models begins with a predata choice of a collection of models likely to be relevant in the field in question, followed by an examination of the reasonability of the data in the light of each model. For those models for which the data seem unreasonable, we have a choice:
> (a) drop them from consideration or
> (b) move them sufficiently close to a smoothed version of the data to make the data reasonable.
>
> Here reasonability is a yes-no decision, not a probability reduction, and the models are thought as challenges, trying to mark the boundaries of reasonability, not to represent likely outcomes. Taking the worst of what remains is a conservative but, in my judgment, reasonable step.

Local Model Uncertainty

An unweighted set description of model uncertainty is also preferred by Lars Hansen and Thomas Sargent, who initiated a broad research programme addressing model uncertainty in macroeconomics (see Hansen and Sargent 2006, for a book-length development of their research plan). In contrast to Tukey, their choice of the unweighted representation is primarily motivated by the difficulty of formulating a sensible prior over a large set of models.

Hansen and Sargent's approach to model uncertainty is an example of the local approach. They assume that by an unspecified search process a policymaker comes to a single approximating model of the economy. Then, the model uncertainty set is built around this model. The set includes all models that are statistically difficult to distinguish from the original approximating model.

More formally, Hansen and Sargent (2006, p. 8) consider a policymaker whose approximating model can be formalized as a Markov process characterized by transition density f ðytjyt_1Þ, where y_t is a state vector at time t. The policymaker's model uncertainty set consists of the Markov processes with transition densities $g(y_t|\, y_{t-1})$, which are difficult to statistically distinguish from the approximating model in the sense that the expected discounted sum of conditional relative entropies of models g with respect to model f is reasonably small:

$$E_g \sum_{t=0}^{\infty} \beta' \int \log\left(\frac{g(z|\, y_t)}{f(z|\, y_t)}\right) g(z|\, y_t) dz \leq \eta.$$

The conditional relative entropy measures the mean information for discrimination between g and f on the basis of a new observation of the state vector, which comes from g (see Kullback and Leibler 1951). What 'reasonably small' means depends on how uncertain the policymaker is about her approximating model. When η is large, the amount of uncertainty may be very large. Hence, the classification of the Hansen–Sargent approach as 'local' does not mean that the uncertainty they address is insignificant. Anderson et al. (2003) relate η to a more transparent concept of detection error probability, which can be used to calibrate η.

Using the relative entropy concept for the formulation of the model uncertainty set is very convenient for design of macroeconomic policy that works well across all models from the set. In an engineering context, Petersen et al. (2000) show how to construct a risk-sensitive control problem which has the same solution as the problem of finding a controller that maximizes the worst possible performance over the set of models subject to the relative entropy constraint.

The risk-sensitive control problem is extensively studied in Whittle (1990). It has a very simple solution, which is a modification of a standard solution of the linear quadratic Gaussian problem. Hansen and Sargent (2006) substantially modify and extend the control methods so that they are applicable to economic problems.

A very important economic setting that calls for an extensive modification of the engineering ideas is that with multiple decision-makers. The rational expectations literature assumes that the economic agents living inside the model and the policymaker who uses the model to formulate her policy agree on the model. The possibility that the agents and the policymaker have doubts about the model calls for a revision of the rational expectations paradigm. Giannoni (2002) is an early example of a study that assumes model uncertainty on behalf of the policymaker but requires the modelled economic agents to know the true model. Hansen and Sargent (2003) consider a situation when the policymaker and the economic agents are uncertain about a common approximating model.

Another example of the local approach to model uncertainty is provided by Schorfheide and Del Negro (2005). In contrast to Hansen and Sargent, they represent model uncertainty about an approximating model by a prior distribution in the model space, centred at the approximating model. To cope with the difficulty of specifying sensible and manageable priors over a vast set of models, they restrict attention to the alternative models that have form of identified vector autoregressions (VARs). The prior density over the alternative models is taken to be proportional to the relative entropy distance between the alternative and the approximating model, which is chosen to be a state-of-the-art dynamic stochastic general equilibrium model.

Using identified VARs for construction of the model uncertainty set potentially raises an extra model uncertainty issue: which identification scheme to use to identify structural shocks in VARs? Different identification schemes cannot be evaluated on the basis of data because the implied identified VARs are observationally equivalent. To take into account uncertainty about the identification schemes, Faust (1998) proposes forming a set of identified VARs so that the corresponding impulse responses look reasonable in the sense that they are consistent with some particularly strong prior beliefs about the effects of structural shocks.

Evolving Model Uncertainty

Schorfheide and Del Negro's (2005) analysis of policy choice under model uncertainty is one of few studies that allows for changes in the model uncertainty depending on the prospective policy choice. Such flexibility comes from their obtaining the joint posterior distribution for the set of possible models and policy parameters. As long as policy parameters are set in the historically observed region, the policymaker can take as his or her model of model uncertainty the posterior distribution over the set of models conditional on the particular parameter values.

Policymakers' perceptions of model uncertainty may depend on many factors beyond particular policy choices. As new data emerge, policymakers learn and adjust their model uncertainty sets. Even more importantly, unforeseen events may substantially change the set of possible models. This fact is at the heart of Keynes's (1939, p. 567) critique of Tinbergen's econometric method: '[The] main prima facie objection to the application of the method of multiple correlation to complex economic problems lies in the apparent lack of any adequate degree of uniformity in the environment.'

An interesting theory of learning under the condition of model uncertainty in a non-stationary environment is Epstein and Schneider (2006). These authors consider a decision-maker who receives a sequence of signals generated by an uncertain model. Some features of the model are constant over time. Those features are represented by a parameter θ, which the decision-maker hopes to learn about, although it is ambiguous initially. Other features may 'vary over time in a way . . . [the decision-maker] does not understand well enough even to theorize about and therefore she does not try to learn about them' (see Epstein and Schneider 2006, p. 3). These features are captured by an assumption that the decision-maker

considers a non-singleton set of data distributions, which are all parameterized by θ but have different structure. Which structure is used to deliver observations may erratically change over time.

Epstein and Schneider show how the set of priors for θ changes over time, and prove that, under certain regulatory conditions, it converges to a distribution assigning probability 1 to a single vector θ* so that the ambiguity about θ is asymptotically resolved. On the contrary, by assumption, the uncertainty associated with the multiplicity of the structures representing the poorly understood factors influencing the dynamics is never resolved or learned about.

Can we form any idea about the nature and the strength of the poorly understood factors? After all, we have to somehow specify a set of distributions representing these factors to analyse model uncertainty. Tetlow (2006) may be a first step in answering this question. He studies the real-time evolution of the principal macroeconomic model of the Federal Reserve Board in the 1996–2003 period. He finds a surprisingly large amount of variation in the model over the period, and shows how the changes in the model were driven by the data and 'the economic issues of the day'.

The literature on model uncertainty is large and rapidly growing. In engineering, the entire field of automatic control is motivated to a large extent by issues of robustness to model uncertainty. Although above we have given an example of engineers' approach to model uncertainty, we have not even scratched upon the surface of the literature. Similarly, many important approaches to model uncertainty in statistics and economics have been left aside. We hope, however, that the reader has gained a general idea about the topic and will find a further discussion in the references provided below.

See Also

- ▶ Ambiguity and Ambiguity Aversion
- ▶ Model Averaging
- ▶ Models
- ▶ Robust Control
- ▶ Specification Problems in Econometrics
- ▶ Uncertainty

Bibliography

Anderson, E., L.P. Hansen, and T.J. Sargent. 2003. A quartet of semigroups for model specification, robustness, prices of risk, and model detection. *Journal of the European Economic Association* 1: 68–123.

Berger, J., and L.M. Berliner. 1986. Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Annals of Statistics* 14: 461–486.

Box, G.E.P., and G.C. Tiao. 1962. A further look at robustness via Bayes's theorem. *Biometrika* 49: 419–432.

Brock, W.A., S.N. Durlauf, and K.D. West. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 2003(1): 235–322.

Brunner, K., and A. Meltzer. 1969. The nature of policy problem. In *Targets and indicators of monetary policy*, ed. K. Brunner and A. Meltzer. San Francisco: Chandler Publishing Company.

Diels, H. 1951. *Die Fragmente der Vorsokratiker*. Berlin: Weidmannsche Verlagsbuchhandlung.

Dow, S.C. 2004. Uncertainty and monetary policy. *Oxford Economic Papers* 56: 539–561.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B* 57: 45–97.

Draper, D., J.S. Hodges, E.E. Leamer, C.N. Morris, and D.B. Rubin. 1987. A research agenda for assessment and propagation of model uncertainty. Report No. 2683-RC. Santa Monica: Rand Corporation.

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistican* 37: 36–49.

Epstein, L.G., and M. Schneider 2006. Learning under ambiguity. Working Paper No. 527, Rochester Center for Economic Research.

Faust, J. 1998. The robustness of identified VAR conclusions about money. *Carnegie-Rochester Conference Series on Public Policy* 49: 207–244.

Fisher, R.A. 1935. *The design of experiments*. Edinburgh: Oliver & Boyd.

Frankel, J.A., and K.E. Rockett. 1988. International macroeconomic policy coordination when policymakers do not agree on the true model. *American Economic Review* 78: 318–340.

Giannoni, M. 2002. Does model uncertainty justify caution? Robust optimal monetary policy in a forward-looking model. *Macroeconomic Dynamics* 6: 111–144.

Gilboa, I., and D. Schmeidler. 1989. Maximin expected utility with non-unique priors. *Journal of Mathematical Economics* 18: 141–153.

Hansen, L.P., and T.J. Sargent. 2003. Robust control of forward-looking models. *Journal of Monetary Economics* 50: 581–604.

Hansen, L.P., and T.J. Sargent. 2006. Robustness. Working paper, New York University.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14: 382–417.

Keynes, J.M. 1921. *A treatise on probability*. London: Macmillan. Repr. for the Royal Economic Society as *Collected Writings*, vol. 8, 1973.
Keynes, J.M. 1939. Professor Tinbergen's method. *Economic Journal* 49: 558–568.
Knight, F.H. 1921. *Risk, uncertainty and profit*. New York: Houghton Mifflin.
Kullback, S., and R.A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22: 79–86.
Kwakernaak, H., and R. Sivan. 1972. *Linear optimal control systems*. New York: Wiley.
McCallum, B.T. 1988. Robustness properties of a rule for monetary policy. *Carnegie-Rochester Conference Series on Public Policy* 29: 173–203.
Morgan, M.S., and M. Morrison. 1999. *Models as mediators*. Cambridge, UK: Cambridge University Press.
Petersen, I.R., M.R. James, and P. Dupuis. 2000. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control* 45: 398–412.
Popper, K.R. 1962. *Conjectures and refutations*. New York/London: Basic Books.
Savage, L.G. 1954. *The foundations of statistics*. New York: Dover Publications.
Schorfheide, F., and M. Del Negro. 2005. *Monetary policy analysis with potentially misspecified models*. Mimeo: University of Pennsylvania.
Tetlow, R.J. 2006. *Real-time model uncertainty in the United States: 'Robust' policies put to the test*. Division of Research and Statistics, Federal Reserve Board: Manuscript.
von zur Muehlen, P. 2001. Activist vs. non-activist monetary policy: Optimal rules under extreme uncertainty. Finance and economics discussion series working paper no. 2, Federal Reserve Board.
Whittle, P. 1990. *Risk-sensitive optimal control*. New York: Wiley.
Zhou, K., J.C. Doyle, and K. Glover. 1996. *Robust and optimal control*. Upper Saddle River: Prentice Hall.

Models

Mary S. Morgan

Abstract

Philosophical analysis of the historical development of modelling, as well as the programmatic statements of the founders of modelling, support three different functions for modelling: for fitting theories to the world; for theorizing; and as instruments of investigation. Rather than versions of data or of theories, models can be understood as complex objects constructed out of many resources that defy simple description. These accounts also suggest a kinship between the ways models work in economics and various kinds of experiment, found most obviously in simulation but equally salient in older traditions of mathematical and statistical modelling.

Keywords

Business cycles; Caricatures; Correspondence rules; Cowles Commission; Design of experiments; Econometrics; Economic man; Edgeworth, F.Y; Experiments and econometrics; Fisher, I; Friedman, M; Frisch, R.A.K; Haavelmo, T; Ideal type; Instrumentalism; Idealization; Inference; Koopmans. T.C; Laboratory experiments in economics; Lucas, R; Macroeconometric models; Matching; Mathematical economics; Mathematics and economics; Metaphor; Methodology of economics; Mill, J.S; Marshall, A; Model design; Models; Model construction; Model functions; Model experiments; National Bureau of Economic Research; Pigou, A.C; Prediction; Probability; Quesnay, F; Random shock models; Shubik, M; Simulation; Slutsky, E; Sutton, J; Statistical inference; Statistics and economics; Tableau économique; Tendency laws; Testing; Tinbergen, J; Weber, M

JEL Classification

B4

Modelling became the dominant methodology of economics during the 20th century.

Yet, despite its ubiquitous usage in modern economics, the term 'model' was introduced relatively recently. In the late 19th century, 'models' were not even a recognized category in discussions about methodology (as for example in Palgrave's *Dictionary of Political Economy* of the 1890s), although a few existed as practical working objects. The effective usage of the term 'model' in economics is associated with the econometrics

movement of the interwar period, a movement whose aim was both to develop and to meld together mathematical and statistical approaches to economics. From this original broad notion, in the 1950s grew separate fields of mathematical economists and econometricians, and both maintained modelling as a central tool of their scientific practice. It became conventional then to think of models in modern economics as either mathematical objects used in economic theory or as econometric objects (involving both statistical and mathematical properties) in empirical work. Historical accounts of models in modern economics may conveniently begin then with this division.

Philosophical commentaries, too, have mostly tended to follow this division, treating the models of economic theory as different kinds of creatures, with different roles, from those which are applied to data. The latter role of models, that of 'fitting theories to the world', is exemplified in the empirical modelling, econometric work and methodological statements of Jan Tinbergen in the 1930s. By contrast, the mathematical models of modern economics are primarily viewed as a way in which economic theory building goes on. This 'modelling as theorizing' view is exemplified in the programmatic pronouncements of Tjalling Koopmans in 1957. A third methodological framework presents 'models as investigative instruments': tools to learn about economic theory or the economic world, a position typified in the late 19th century and early 20th century work of Irving Fisher, who might be seen as another of the founders of modelling in economics.

This article covers the historical emergence and roles of models in economics according to these three different methodological accounts, and discusses how these approaches fit into the modern science of economics.

Modelling as Fitting Theories to the World

Although a vibrant econometrics community developed in the two decades up to the 1920s, its products (regressions of demand, statistical accounts of business cycles and so forth) were presented as direct descriptions of the underlying economic relations, rather than as models put forward tentatively to represent them (see Morgan 1990). The difference is a subtle one, but illuminated by Philippe Le Gall's (2007) use of the term 'natural econometrician' for those 19th century economists who believed, in parallel to the natural sciences, that the laws that governed the economy were written in mathematics, and clever manipulation of statistical data (without, it must be said, much in the way of analytical techniques) would reveal these laws.

Into this descriptive statistical framework, Jan Tinbergen not only introduced the term 'model' in 1935 (see Boumans 1993) but he was also responsible – along with Ragnar Frisch – for the development of such joint mathematical–statistical objects in the econometrics of the 1930s. (Prior to this, the rare use of the term 'model' typically referred to physical object models as Boltzmann defined them in 1911. Paul Ehrenfest is the probable source of a broadening in scope of the term to include mathematical models, and Tinbergen was his assistant during the mid-1920s; see Boumans 2005, ch. 2.) Frisch in 1933 had developed – in the context of business cycle research – the notion of a 'macro-dynamic scheme': a three-equation model with random errors. He even simulated it to show that it could reproduce the generic characteristics of time-series data of his time. But it was Tinbergen who developed Frisch's design into an econometric model – a model that could be fitted to real data from the economy. As is well known, he built the first generation of macroeconometric models (see Tinbergen 1937; 1939; and Bodkin et al. 1991), and in doing so he made explicit the notion of a model as a vehicle for bridging the gap between theories of the business cycle and specific (time and place) statistical data of the cycle, as Morgan (1990, ch. 4) argues. To appreciate the task, it needs to be remembered that most existing theories of the cycle were expressed verbally, and the nascent mathematical theories of the cycle were too small and simplified to represent the characteristics of real cycles, so even building up a system of equations from these theories was a considerable task. The data played a role, too, in

deciding the time sequence of the relations and which variables should be included or omitted, for both these elements were determined in the statistical work. In other words, Tinbergen created a set of usable mathematical–statistical relations which both incorporated theoretical ideas about how the economy worked and represented empirically the different parts of the economy. Having fitted theories and data together in the format of the econometric model, he then used the model to test the viability of various theories of the cycle, to explain events in the economy, and to run the model forward with different policy options relevant for the Great Depression years – all this in the pre-computer age using hand calculators! This 'new practice' of models, as Boumans (2005) terms it, involved a creative building of mathematical economic theory in relation to the statistical data of the economic world and of craft skill in using those models. For both Frisch and Tinbergen, modelling was a project to explain how the economic world worked.

The next stage in the history may be marked by Trygve Haavelmo's famous blueprint for econometrics of 1944 which brought another subtle change of focus to the task of econometric modelling. He suggested that econometrics ought to be concerned, not with a process of matching theory and data in an iterative process, but with finding the correct model for the observed data using probability reasoning (see Morgan 1990, ch. 8). He effectively introduced into econometrics not only the notion of the theoretical model (the mathematical model derived from a priori theory) but also that of the 'true' (but unknown) model: 'the 'true' mechanism under which the data considered are being produced' (Haavelmo 1944, p. 49). Yet he was by no means a 'natural econometrician' (in Le Gall's sense for the 19th century), arguing of models of economic behaviour that 'whatever be the 'explanations' [of economic phenomena] we prefer, it is not to be forgotten that they are all our own artificial inventions in a search for an understanding of real life; they are not hidden truths to be "discovered"' (Haavelmo 1944, p. 3). Though he urged that a well- fitting econometric model (a theory which fits the data well) might not be the 'true' model, nevertheless, his blueprint probability approach was destined to alter the accepted task of econometrics. The Cowles Commission approach that followed (whose contributions are analysed by Qin 1993, and Epstein 1987) stressed the use of the correct methods of identifying and estimating the theoretically derived complete structural model as the means to discover that true model. The 'strong apriorism' of their approach to econometrics, in which theory proposes the model and the data dispose (or not) of these hypotheses, sparked the famous 'measurement without theory' debate with the more empiricist branch of the field over how to do econometrics in the late 1940s.

It is tempting to see Haavelmo's provision of a philosophical basis for econometrics as paving the way for a post-1950 division of labour in the use of models – namely, the economists provide mathematical models from economic theory, and the job of the econometrician is to use statistics for model estimation and theory testing. To some extent this division of labour is borne out, for it is in this period that a much clearer distinction emerges between theoretical and applied economics (as seen in Backhouse 1998). However, despite the rhetoric of post-1950 econometrics which talks of 'confronting theory with data', or 'applying theory to data', from the point of view of econometric modelling the practical division is not nearly so clear-cut. There are several reasons. First, it remains a prosaic but generally valid comment that theory rarely provides all the resources needed to make models that can be immediately applied to the data from the world. This is precisely why econometric models have featured as a necessary intermediary, a matching device, between them. Second, this matching process of fitting theories to the world is done with many different purposes – to test theories, to measure relations, to explain events, and so on – each needing different resources from theory and with different criteria. Third, there are no general scientific rules for modelling. There have been fierce arguments within the econometrics community in recent decades over various scientific principles for modelling (and associated criteria): whether models should be theory driven or data driven; whether the modelling process should be simple

to general or general to specific; and so forth (see Pagan 1987; Heckman 2000). Regardless of which principles are followed, the creative element is still very much evident wherever applied econometric modelling occurs, whether such modelling is at the pattern-seeking end of the spectrum or theory-led modelling, and whether the field is macro- or micro-econometrics.

A more recent shift in focus, particularly in the macroeconometric field and associated with Robert Lucas, entails giving up on the aim of using theory to make models that represent the true general structure as a way to uncover that structure. As he wrote:

> A 'theory' is not a collection of assertions about the behavior of the actual economy but rather an explicit set of instructions for building a parallel or analogue system – a mechanical, imitation economy. A 'good' model, from this point of view, will not be exactly more 'real' than a poor one, but will provide better imitations. (Lucas 1980, p. 697)

This move changes the relation between models and theory, for now the task of theory is to produce models as analogues of the world, rather than to use them to explain the behaviour of the world (see Boumans 1997). At the same time, it shifts the focus of 'fitting': the aim is no longer to fit theory to the world but to fit the model to the world in the particular sense of being able to imitate certain sorts of data characteristics.

Another recent account, developed this time in micro-econometrics by John Sutton (2000), validates itself in relation to the earlier econometric agenda held by Frisch and Tinbergen, for, like those early pioneers, he thinks of models not as devices for the discovery of the true general model as in the Cowles Commission interpretation of Haavelmo's project, nor as mathematical machines that imitate the world as in Lucas's account, but as investigative devices for finding out about the world. In Sutton's view, the economic world produces reasonably stable regularities or variability only within a class of cases, not across all cases; thus, looking for a general model is too ambitious. The aim of modelling is to describe the economic mechanisms that produce the data characteristics that are shared within a subset of all cases and so explain the regularities observed within that subclass. Sutton describes this as a 'class of models' approach. Once again, models appear as an intermediary device between theory and data, but this time function to sort out like cases in the world and so offer explanations for their characteristic behaviour.

Models apparently play a critical epistemological role in econometrics – but there are different ways of characterising this. Econometrics can be seen as fulfilling the function of laboratory experiments in some other sciences – a claim that lies implicit in Haavelmo's discussion of the data of economics as being the result of passive observation of nature's experiments and explicit in his discussion of econometric modelling as designing experiments (see Haavelmo 1944, chs. 1 and 2). His conceptualization of econometrics appeals to the importance of probability and statistical reasoning as the bases for both model design and statistical inference: models have to be designed to match data that could be observed, and be framed in probability terms. The 'design of experiments' notion requires the econometrician to think about the fitting problem, while the probability set-up gives rules for inferences from the model experiment, ones that are in fact much better specified than those for laboratory experiments in most sciences. Thus Haavelmo's blueprint explicitly buys into a tradition of statistical thinking as a valid mode of scientific reasoning, but reinterprets it as a form of experimental work.

A more recent characterisation of the epistemological function of models in econometrics is to understand them as instruments of observation and measurement that enable economists to identify stable phenomena in the world of economic activity. Kevin Hoover's account of 'econometrics as observation' describes 'econometric calculations' as 'the economist's telescope' (1994, p. 74) where rules for focusing the telescope come from statistical theory and where economic theory, and the purpose engaged in, guide the observation process. Marcel Boumans (2005) understands models as the primary instrument in this process, without which the economists could not 'model the world to number'. Rather than a means of observation, he portrays models as complex scientific instruments that generate the

numbers for those economic objects, concepts and relations that cannot be observed directly and that are not yet measured. Like Haavelmo, Boumans's account of model work invokes a careful design of experiments, but he provides a more concrete discussion of how econometric modelling provides measurement structures to deal with *ceteris paribus* clauses; how statistical and other criteria provide ways of assessing the reliability of model instruments (via calibration, filtering and so forth); and how precision and rigour are obtained in the measurement process.

Neither Boumans nor Hoover is instrumentalist about models in the sense that has come to be associated with Milton Friedman's 1953 argument that models need be efficient only for prediction, not for explanation. (Friedman's essay has been much argued over, and interpretations of this particular point vary; see particularly Hirsch and De Marchi 1990; instrumentalism and operationalism; and Mäki 2007.) Nor are they operationists in the Bridgmanian sense (that informed, for example, Paul Samuelson's early work in economics; see Bridgeman 1927), namely, that a concept is defined by its measuring process (such as an econometric model). Both Hoover and Boumans might be termed 'sophisticated instrumentalists' for they regard econometric calculations or models as cleverly designed instruments for observing and measuring the relations of economics, and so understanding and explaining, the world.

Modelling as Theorizing

The term 'model' had rarely been used in economics before the 1930s, even though things we would now label 'models' had been developed and used for theorizing before then. We can certainly recognize some earlier examples of modelling in the late 19th century; for example, we can happily denote the Edgeworth–Bowley box diagram, and Alfred Marshall's trade diagrams and supply–demand scissor diagrams as models. These examples signal that modelling was an unrecognized element in the mathematizing process of that earlier period (see Morgan 2008). Yet it was only after the 1950s that modelling became a widely recognized way of using mathematics in economics and became one of the dominant forms of economic theorizing. Whereas the establishment of the statistical–econometric notion is associated with Tinbergen, the mathematical–theorizing one may be associated with another Dutch econometrician, Tjalling Koopmans, whose account, given in a set of three essays in 1957, is widely understood as a paradigmatic statement of the modelling approach of modern mathematical economics. Koopmans had developed Tinbergen's earlier ideas about modelling to fit with contemporary discussions of the role of mathematics in economics in the 1940s and 1950s and with the formal mathematical idea of a model at that time. As such, his statement fits into a broader history of mathematics and economics treated particularly in Weintraub (2002) and Ingrao and Israel (1987).

Koopmans defined an economic theory as a set of postulates with which we reason in order to work out and make explicit the otherwise implicit effects of the set of postulates taken together: a reasoning practice that apparently involves models. For Koopmans, this reasoning was an important part of theorizing since these implications are not self-evident, nor is any particular set of postulates necessarily fruitful. His portrayal of 'Economic Theory as a Sequence of Models' (to quote his 1957, p. 142, section title) is presented as his answer to the ongoing argument of his day about the status of the assumptions and the predictions of economics in which he explicitly defined the role of models almost as an aside:

> neither are the postulates of economic theory entirely self-evident [as Robbins had argued in 1932], nor are the implications of various sets of postulates readily tested by observation [as Friedman had argued in 1953]. In this situation, it is desirable that we arrange and record our logical deductions in such a manner that any particular conclusion or observationally refutable implication can be traced to the postulates on which it rests . . . Considerations of this order suggest that we look upon economic theory as a sequence of conceptional *models* that seek to express in simplified form different aspects of an always more complicated reality. At first these aspects are formalized as much as feasible in isolation, then in combinations of increasing realism. (Koopmans 1957, p. 142)

Koopmans suggests, then, that models are an essential element in theorizing, and that their role

comes in their sequenced ability to express different and combined aspects of a simplified reality. But his projection that such a sequence of models would represent 'combinations of increasing realism' seems not to have been borne out. While tractability suggests that increasing realisticness in some aspects will have to be traded off against simplification in others, the history of modelling suggests that model sequences are more often driven by changes in problems, in questions, and in the mathematical tools available. This last was a possibility that Koopmans himself discusses in the context of the move from arithmetical to diagrammatic to algebraic forms of theorizing. And, as just noted with Lucas, some modern modelling no longer aims to represent the world as it is, but to develop artificial systems that mimic outputs from the world.

There are various ways of characterizing the use of mathematical models in economic theory. For Daniel Hausman, the connection of models with concept formation is both more explicit and more important than Koopmans suggests, for economic modelling is where theory development goes on:

> A theory must identify regularities in the world. But science does not proceed primarily by spotting correlations among various known properties of things. An absolutely crucial step is constructing new concepts – new ways of classifying and describing phenomena. Much of scientific theorizing consists of developing and thinking about such new concepts, relating them to other concepts and exploring their implications.
>
> This kind of endeavor is particularly prominent in economics, where theorists devote a great deal of effort to exploring the implications of perfect rationality, perfect information, and perfect competition. These explorations, which are separate from questions of application and assessment, are, I believe, what economists (but *not* econometricians) call 'models'. (Hausman 1984, p. 13)

Nowadays, the explorations would be into bounded rationality, imperfect information and imperfect competition: the agenda has moved on, but the mode of theorizing via modelling remains the same. Hausman's attention to the role of models in conceptual innovation is given credence and depth in his own analysis of Samuelson's 'overlapping generations' model, a story about creative exploration in the theoretical realm. The Edgeworth Box history (see Humphrey 1996, and Morgan 2004a) provides another good example of the way modelling is associated with new concepts and descriptions – it is after all where indifference curves, contract curves and so forth were first introduced.

The development of the supply and demand diagram we find in Marshall's *Principles* (1890) exemplifies Hausman's claims. It is not just that Marshall's diagrams describe in new ways some older ideas about the phenomena of supply and demand that go way back in the purely verbal literatures of economics, but that in his hands these curves are fashioned to represent various kinds of markets and relations, resulting in new concepts and classification of types of supply or demand at a level that sits between any general theory and one-off cases (see Morgan 2002). It is this function of modelling as a classification device that Sutton (2000) reprises in a different form in his 'class of models' work on industrial competition (discussed above). And, historically between these two economists, we can situate, as just one example, the work by Martin Shubik (1959) who used game theoretic models to classify kinds of competition and industry structure according to the kind of game that most matches the economic situation involved.

Hausman is keen to make his account of the methodology of economics not only fit to the practice of modern economics, but philosophically sensible, so he separates the activity of modelling from the more general assertions and truth claims of theories. At first sight this strict separation may look curious to economists who often talk of 'testing models' rather than theories, and do not bother to pull apart the categories of theories and models in their everyday scientific work. This conflation may occur because, as Hausman suggests, 'Models are not themselves empirical applications, but they have the same structure' (Hausman 1992, p. 80). Having the same structure might enable empirical application by econometricians, though this is not how economists mostly use mathematical models in arguing about the world: rather, they are more often linked to the world in a much more casual fashion.

Indeed, 'casual application' is exactly the term used by Alan Gibbard and Hal Varian to describe how mathematical models are applied 'to explain aspects of the world that can be noticed or conjectured without explicit techniques of measurement' (1978, p. 672). In their view, mathematical models are designed only to *approximate* the world, and, unlike econometric models which go through a serious process of fitting to the world, they are casually connected to the world by 'stories' which interpret the terms in the model to elements in the world. But they stress that such applications of models do not pertain to particular situations or things in the world. In contrast, Hausman (1990) argues that economists do often use their models in this way to discuss particular real world events, and they use narratives to fill in the descriptions given in the model in order to provide explanations of those events in the world. Morgan (2001, 2007) takes a stronger position with regard to these stories, suggesting that they form an integral part of the application of models to the world – both in general and for particular cases – and equally form an essential part of the identity of the model. Steven Rappaport (1998), like Hausman, finds mathematical models to be quite stretchable in function: in conceptual work, in normative work (for example in discussions of policy), and in heuristic explanatory work. However, in other respects Rappaport's account of models and their function contrast with Hausman's and with Morgan's, for he portrays models as 'mini-theories' within a research programme that function in counterfactual format: that is, their function is to provide accounts of what might happen if the model were a true description of the world.

These accounts of how mathematical models connect to the world all suggest a dependence on cognitive, intuitive or informal elements of economists' theorizing with respect to the world, in strong contrast to the statistical and economic criteria that attend the way econometricians use models to fit theories to the world. On the other hand, mathematical models appear to fulfil a wider variety of functions ranging from devices for new concept formation and classificatory work in theorizing to inference devices that purport to give explanations of general or particular events. Policy usage often involves mathematical models for analysis of policy interventions and for mechanism design purposes – as, for example, in the design of auctions. So far there is little historical or reflective philosophical literature on this side of model work (though see Guala 2001). By contrast, there is a considerable reflective literature on the policy activities associated with empirical or econometric models (see examples and references in Den Butter and Morgan 2000).

Models as Investigative Instruments

We have already seen various ways in which models are understood as investigative devices. In the commentaries on econometrics, we found models portrayed as tools or instruments of observation and measurement, and in the early econometric work models were also understood as tools to help explain the world. The idea of models as instruments is also present in the mathematical modelling literature, but is associated with a more active sense of investigation. Irving Fisher, for his thesis, physically built a three good, three consumer, hydraulic analogue general equilibrium model:

> The mechanism just described is the physical analogue of the ideal economic market. The elements which contribute to the determination of prices are represented each with its appropriate role and open to the scrutiny of the eye. We are thus enabled not only to obtain a clear and analytical *picture* of the interdependence of the many elements in the causation of prices, but also to employ the mechanism as an instrument of investigation and by it, study some complicated variations which could scarcely be successfully followed without its aid. (Fisher 1892, p. 44)

This chimes well with the commentary from Scott Gordon, who, from his historical and philosophical analysis of economics, claims that 'the purpose of any model is to serve as a tool or instrument of scientific investigation' (1991, p. 108).

The notion of tools in economics has not been well-developed. Arthur Pigou (1929) introduced the distinction between 'tool makers' and 'tool users', labelling Francis Edgeworth as a maker

of tools, and Marshall as both a maker and user. For Pigou, the term 'tools' referred not to processes of induction as opposed to deduction, or even to the mathematical as opposed to the literary method, but to something he referred to as a 'wider' analytical movement involving specific statistical and mathematical techniques or 'machinery' (such as the method of analysis of demand and supply). It was in following him that Joan Robinson (1933), in oft-quoted comments, wrote about the 'tool-box of economics' which she presented as consisting of 'assumptions' (theory) and 'geometry' (methods) though we might more naturally think of these combining to form models. Koopmans (1957), too, wrote about tools, referring not only to numerical examples and diagrammatic representations, but also to formal mathematics, computing techniques, input–output analysis and so forth, thus (for our time) mixing up methods or modes of analysis (ones we associate now with modelling) and kinds of models. Yet there is a striking similarity between the way Fisher referred to and used his physical hydraulic model and the way modern economists use their equivalent mathematical models of modern economics as tools of investigation. Both seem to be well covered by the notions of tool using that Pigou introduced.

Indeed, attention to the functions of models has emphasized that much of the classifying and conceptual development work of theorizing discussed in the previous section occurs not so much in building mathematical models as in using them. For example, the models developed by Hicks, Samuelson, Meade and others in the late 1930s based on Keynes's *General Theory* were used to explore, develop and understand that theory in ways that involved substantive conceptual and classifying work of their own (see Darity and Young 1995). In deriving solutions to theoretical problems, or in exploring the limits of behaviour implied by the theoretical relations represented in the models, and in applying their models to think about problems of the economic world represented in the model, those economists used their models as instruments of investigation. These investigations appear as glorified thought experiments, too complicated to do in the mind and so requiring a representation of the case or system in the form of the model and associated mathematical modes of reasoning about it. In Fisher's case, he had a material object to experiment with. Mathematical models in economics also typically provide such internal resources for experimental manipulation. Morgan (2002) argues the case for regarding mathematical modelling activity as experimental work on mathematical models in parallel with statistical experiments practised on econometric models. But whereas we have well-grounded statistical rules for making inferences from econometric experiments, the application of mathematical models to the world (or inferences from such model experiments) is more casual or approximate, as we have already seen.

This notion that mathematical modelling work is a form of experimental activity is most evident in the founding literature on simulation in economics around 1960 (surveyed at the time by Shubik 1960a, b). In some other fields of science, simulation has been introduced primarily as a method of numerical, rather than analytical, solution. But in economics, simulation has been more usually presented and used as a process of experiment on models, a process that effectively investigates in a systematic manner the full range of behaviours of the system or the actors portrayed in the model. There were isolated examples of simulation earlier in the history of economics – most particularly Tinbergen's 1936 simulation of his macroeconometric model, Paul Samuelson's (1939) simulation of a little Keynesian mathematical system and Eugen Slutsky's (1927) famous random shock models that mimicked business cycles. The possibilities of simulation were then explored more effectively during 1950s and 1960s Cold War activities that brought the social sciences and mathematics together.

The birth of simulation in economics has usually been attributed to Herbert Simon, but equally important were concurrent developments connected with other pioneers, particularly Frank and Irma Adelman, Martin Shubik and Guy Orcutt (see Morgan 2004b). Simon's simulation projects in economics involved, for example, programming computers to imitate decisions and choices in the same way that investment bankers made those decisions and choices, that is, on the

same information and by the same processes of comparison and assessment (see Clarkson and Simon 1960). The Adelmans's work was particularly important in the development of simulation methods in econometrics following the lead of Tinbergen's earlier work (see also Duesenberry et al. 1960), while in economics at that time simulations involved both 'game playing', meaning experiments in which people role-played making economic decisions where the model simulated the environment and all the interest was in the behaviour of the people (for example, managers making decisions), and mathematical model simulations in which the behaviour was taken as given (for example, rational economic behaviour) and the environment varied to see how that altered the outcomes projected by the model. (This broad category of simulations around 1960 thus included some things we would now label experiments.) Shubik was involved in many of these different types of simulations ranging from game-playing experiments, to business games, to model experiments. Orcutt (1960) meanwhile pioneered the method of microsimulation, in which he constructed a representative virtual sample of the population, endowed the sample individuals with characteristics of the real population, and then simulated their behaviour through time to explore the characteristics of the aggregate system as well as the individual parts. This is complicated model-experimental work that was possible only with the new-found computing power of that day. All these economists significantly extended the ways in which models worked as instruments of investigation via different forms of experimental activity in which each 'run' of the model provided a slightly different experiment with the model. Simulation, since its introduction into economics, has been characterized as a form of experiment with models that aims at mimicking a variety of different economic behaviours, at different levels and in different ways.

Model Construction

Model making (as opposed to formal or informal definitions of models) has been a fertile ground for philosophical commentators on economics who have presented it as a process of 'idealization', a term that covers a range of things including abstraction, simplification and isolation (see Hamminga and De Marchi 1994). This general idea goes back to the 'ideal type' concept defined by Max Weber (1904, 1913) for the social sciences. His discussion included notions of the ideal type of individual economic behaviour and the ideal type notion of a market. Certainly it is easy to see the late 19th century portrait of economic man as ideal type, divorced from all but his pure economic motivations without any deeper psychology. The term 'idealization' suggests that models are arrived at by processes of *abstracting* to the level of ideas or concepts; of *simplifying* the case or system treated by omitting irrelevant or negligible influences; of *isolating* the elements that are really thought to be important by *ceteris paribus* clauses; and so forth (see Morgan 2006). These processes can be understood as working on theories (for example, moving from a full equilibrium account down to a single particular market) or as starting with the complicated world and isolating a small part of it for model representation. Leszek Nowak (for example, 1994) presents a rather general analysis in which 'idealization' takes one from the world to theory and 'concretization' from theory to the world in two rather seamless parallel processes. This account known as the 'Poznań approach' (named after the University that hosted its development: see Hamminga 1998), was formulated for Marxian economics, but might well be applied more generally. Two other commentators particularly associated with questions of idealization in economic modelling are Nancy Cartwright and Uskali Mäki. Cartwright (1989) is interested in what has been called 'causal' idealization, that is, in isolating the causal capacities that actually work in the world. She associates this aim both with how econometric modelling works and with Millian tendencies (the account of tendency laws in economics provided by John Stuart Mill in the mid-19th century). Mäki (1992) is more interested in 'construct' idealization, that is, in how economic theorizing goes on by constructing versions of theory with more or less scope along different

dimensions of isolation. (The distinction between construct and causal idealization used here is due to McMullin 1985.) We can find both these kinds of process going on in the history of model making. Von Thünen's (1826) construction of his diagrammatic model of an 'isolated state' provides a clear example of model- making by isolating the factors that determine farm profitability. His isolations can be interpreted as creating a theoretical model (that is, he constructed an idealized model) but he was also interested in getting at real causes for he fitted this model to his own farm's statistical data (that is, he isolated the causes, using informal econometric procedures).

Idealization itself may involve not just simplifications or isolations but the addition of false elements. Max Weber (1904) discusses how ideal types present certain features in an exaggerated form, not just by accentuating those features left by the omission of others but as a strategy to present the most ideal form of the type. This notion of exaggeration comes up again in Gibbard and Varian's (1978) notion of caricature modelling in economics, where the exaggeration is designed to enable the economist to investigate the robustness of the model (the virtue that Friedman had, of course, earlier associated with the use of unrealistic assumptions). But if we interpret this caricaturing process to involve not just an extreme degree of exaggeration but the addition of features, then we have an idealization of a qualitatively different kind from those that come from methods of isolation or simplification. For example, Frank Knight's (1921) assumption of perfect information involves adding a feature to the portrait of economic man; the assumption can be specified in different ways, each creates a different model. Caricature models are not to be confused with the artificial constructions of Lucas's models, which are not derived by idealization from either theory or the world. Idealizations, even in caricaturing form, are still understood as representations of the system or man's behaviour (however unrealistic or positively false these might be) whereas the artificial world models do not seek to represent the system or agent's behaviour – rather, the aim is to mimic the output of such systems or behaviour. In imitating the system outputs, one might of course argue that representational power is sought at a different point.

In economics itself, as opposed to in the analyses of commentators, these processes of model making may all be going on together at the same time. That is, models may be constructed to represent the idealized versions of grander theories, be abstracted from the particularities of economic life, and provide simplifications of the more complicated world. These features are all at play in François Quesnay's famous 18th century *Tableau économique*, a construction that may be regarded as the general ancestor of models in economics. But that model makes a telling example, for as a construction it is only in part a derivation or isolation from a general set of ideas or theory, only in part a simplification of the relations in the world or abstraction into a more conceptual framework. It does not seem to be derived entirely from theory, nor does it appear as a description of his contemporary data. Yet while it does embody elements of all these things, it is also a construction of its own (see Charles 2004). Quesnay moulded these elements together to create a wonderful table-cum-picture that represents the French economy of his day, one that few later economists can understand easily (at least without translating it into a different form, which of course changes its meaning and working).

This interpretation of Quesnay's modelling assumes that models are neither just derived from theory nor solely built up from data, for they typically involve bits of both and oftentimes other things as well, such as metaphors, imported mathematical forms, and so on. The notion that econometric models are constructed from both theoretical relations and statistical elements is probably not that contentious. The mixture of elements is also obvious in a case like the Phillips–Newlyn model, a real hydraulic machine in which red water, representing the various aggregate stocks and flows of the economy, circulated around the machine and sometime spilt into the lecture room (see Leeson 2000; Boumans and Morgan 2004). But these mixtures are equally characteristic in mathematical models, according to the case work account of model building by

Boumans (1999), who argues that we should think of model making as like cooking new recipes, in which mathematics provides the means of integrating such several, sometimes disparate, elements into new models. This account of model construction goes against much traditional philosophizing, even by economists, about model making. Yet more recently economists have begun to write about their modelling work as a much more ad hoc activity in which past practices, new intuitions and even speculations guide their model making (see, for example, Krugman 1993; Sugden 2000).

Understanding model making according to Boumans's recipe-making account suggests that models – by construction – are partially independent of both theory and the world (or its data), and this accounts for their apparently autonomous existence as working objects in modern economics. This construction account is part of the 'models as mediators' view of the role of models, which analyses their use as investigative instruments (see Morrison and Morgan 1999). According to this account, models can function in this autonomous in-between way because of their construction. However, the possibility of learning from using models depends on another element in their construction, namely, that models are devices made to represent in some way or form something in our economic theories or in the economic world or both at once. It is this representing quality, built in at the construction stage, which makes it possible to use a model not just as an instrument of prediction but as an investigative instrument to learn something about the world or the theory which it represents. This account can apply even to the artificial world models proposed by Lucas which are constructed not to represent the workings of the system but the outputs of the system, though here the modellers' ambitions to learn from the modelling in order to understand the economic system and explain the outcome phenomena that they mimic seems somewhat reduced.

This recent recipe account of model-making stands in marked contrast to accounts of how model making goes on according to those mid-20th century commentators discussed earlier. Recall that Koopmans had labelled mathematical models as 'defined by a set of postulates' where the full set of postulates form the theory – a definition consistent with the then current axiomatic approach to theories. In econometrics, the Cowles Commission presented econometric models as being derived – directly given in some sense – from a priori theory. Indeed, it was the basis of their position in the 'measurement without theory' debate that econometrics needed models that were clearly versions of theories to get anywhere at all, against the data-derived models of the National Bureau of Economic Research (NBER) that they decried as unscientific. Another description that fits the philosophical inclinations of the mid-20th century, but is more model-oriented, was given by Friedman, who defined a theory as consisting of two parts: 'a conceptual world or abstract model simpler than the "real world" and containing only the forces that the hypothesis [theory] asserts to be important' and a second part defining the 'class of phenomena for which the "model" can be taken to be an adequate representation of the "real world"' along with the correspondence rules linking the model terms and the phenomena (Friedman 1953, p. 24). Friedman here neatly depicts the model as both a version of theory and at the same time a representation of the real world, yet the correspondence rules are by no means unproblematic. While one could argue that the main work of econometrics has been to develop both the theory and practices of such correspondence rules for models, for mathematical models, in contrast, methodological accounts have often foundered on how such correspondence criteria might be formulated. Despite the long shadow of these rather formal mid-20th century definitions, it is in keeping with our observations about how models are used in modern economic science that they may now be understood as autonomous working objects, rather than as either proto-theories or versions of data.

Conclusion

There is more that might be said, and that remains to be researched, about the philosophy of

modelling, for example about the nature of reasoning with mathematical models; about the role of mathematical models within the design of classroom/laboratory experiments in economics; about the use of models in policy advice and intervention; and about the absence of formal criteria for working with mathematical models that are equivalent to the statistical criteria associated with econometric model work. There is also much to be done in filling in the skeletal history of modelling offered here: in separating the history of modelling from both the history of mathematical economics and the history of econometrics; in demarcating the historical range of scope of modelling; and in discerning why and how the method took hold. Nevertheless, the basic trajectory of the history is clear: modelling becoming defined as a mode of reasoning and working for economics in the 1930s, it was developed and used in various ways in the 1940s and 1950s, setting the scene for modelling to become a dominant methodology in the latter part of the century. And once defined, we can look back and recognize earlier prototypes for such a method going back to Quesnay in the 18th century. When we so look back, and consider the scientific world view that we have lost in economics by adopting modelling as one of our favoured methods of doing economics, what stands out is that the science is a radically different one. No longer do economists believe and enquire into a few grand governing laws, nor even propose wide-ranging general theories – rather, economics has become a science of many different and particular models.

See Also

- ▶ Econometrics
- ▶ Edgeworth, Francis Ysidro (1845–1926)
- ▶ Fisher, Irving (1867–1947)
- ▶ Instrumentalism and operationalism
- ▶ Koopmans, Tjalling Charles (1910–1985)
- ▶ Mathematics and economics
- ▶ Methodology of economics
- ▶ Tinbergen, Jan (1903–1994)

Bibliography

Adelman, I., and F.L. Adelman. 1959. The dynamic properties of the Klein–Goldberger model. *Econometrica* 27: 596–625.

Backhouse, R.E. 1998. The transformation of U.S. economics, 1920–1960. In *From interwar pluralism to postwar neoclassicism*, Annual Supplement to *History of Political Economy*, vol. 30, ed. M.S. Morgan and M. Rutherford. Durham: Duke University Press.

Bodkin, R.G., L.R. Klein, and K. Marwah. 1991. *A history of macroeconometric model-building*. Aldershot: Elgar.

Boltzmann, L. 1911. Models. In *Encyclopaedia britannica*, 11th ed. Cambridge: Cambridge University Press.

Boumans, M. 1993. Paul Ehrenfest and Jan Tinbergen: A case of limited physics transfer. In *Non-natural social science: Reflecting on the enterprise of more heat than light*, ed. N. De Marchi. Durham: Duke University Press.

Boumans, M. 1997. Lucas and artificial worlds. In *New economics and its history*, ed. J.B. Davis. Durham: Duke University Press.

Boumans, M. 1999. Built-in justification. In *Models as mediators*, ed. M.S. Morgan and M. Morrison. Cambridge: Cambridge University Press.

Boumans, M. 2005. *How economists model the world to numbers*. London: Routledge.

Boumans, M., and M.S. Morgan. 2001. *Ceteris paribus* conditions: Materiality and the application of economic theories. *Journal of Economic Methodology* 8: 11–26.

Boumans, M., and M.S. Morgan. 2004. Secrets hidden by two-dimensionality: The economy as a hydraulic machine. In *Models: The third dimension of science*, ed. S. de Chadarevian and N. Hopwood. Stanford: Stanford University Press.

Bridgeman, P. 1927. *The logic of modern physics*. New York: Macmillan.

Cartwright, N. 1989. *Nature's capacities and their measurement*. Oxford: Clarendon Press.

Charles, L. 2004. The Tableau économique as rational recreation. *History of Political Economy* 36: 445–474.

Clarkson, G.P.E., and H.A. Simon. 1960. Simulation of individual and group behaviour. *American Economic Review* 50: 920–932.

Darity, W., and W. Young. 1995. IS–LM: An inquest. *History of Political Economy* 27: 1–41.

Den Butter, F., and M.S. Morgan. 2000. *Empirical models and policy making: Interaction and institutions*. London: Routledge.

Duesenberry, J.S., O. Eckstein, and G. Fromm. 1960. A simulation of the United States economy in recession. *Econometrica* 28: 749–809.

Epstein, R.J. 1987. *A history of econometrics*. Amsterdam: North-Holland.

Fisher, I. 1892. Mathematical investigations in the theory of value and prices. Yale University thesis, repr. New Haven: Yale University Press, 1925.

Friedman, M. 1953. *Essays in positive economics*. Chicago: University of Chicago Press.
Frisch, R. 1933. Propagation and impulse problems in dynamic economics. In *Economic essays in honour of Gustav Cassel*. London: Allen & Unwin.
Gibbard, A., and H.R. Varian. 1978. Economic models. *Journal of Philosophy* 75: 664–677.
Gordon, S. 1991. *The history and philosophy of social science*. New York: Routledge.
Guala, F. 2001. Building economic machines: The FCC auctions. *Studies in History and Philosophy of Science* 32: 453–477.
Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement): iii–iv. 1–115.
Hamminga, B. 1998. Poznań approach. In *Handbook of economic methodology*, ed. J.B. Davis, D. Wade Hands, and U. Mäki. Cheltenham: Edward Elgar.
Hamminga, B., and N. De Marchi (eds.). 1994. *Idealization in economics*. Amsterdam: Rodopi.
Hausman, D.M. (ed.). 1984. *The philosophy of economics: An anthology*. Cambridge: Cambridge University Press.
Hausman, D.M. 1990. Supply and demand explanations and their *ceteris paribus* clauses. *Review of Political Economy* 2: 168–187.
Hausman, D.M. 1992. *The inexact and separate science of economics*. Cambridge: Cambridge University Press.
Heckman, J. 2000. Causal parameters and policy analysis in economics: A twentieth century retrospective. *Quarterly Journal of Economics* 115: 45–97.
Hirsch, A., and N. De Marchi. 1990. *Milton friedman: Economics in theory and practice*. New York: Harvester Wheatsheaf.
Hoover, K.D. 1994. Econometrics as observation: The Lucas critique and the nature of econometric inference. *Journal of Economic Methodology* 1: 65–80.
Humphrey, T.M. 1996. The early history of the box diagram. *Federal Reserve Board of Richmond Economic Review* 82(1): 37–75.
Ingrao, B., and G. Israel. 1987. *The invisible hand: Economic equilibrium in the history of science*. Trans. I. Cambridge, MA: MIT Press McGilvray, 1990.
Knight, F.H. 1921. *Risk, uncertainty and profit*. Boston: Houghton Mifflin.
Koopmans, T. 1957. *Three essays on the state of economic science*. New York: McGraw Hill.
Krugman, P. 1993. How I work. *American Economist* 37(2): 25–31.
Le Gall, P. 2007. *A history of econometrics in France: From nature to models*. London: Routledge.
Leeson, R. 2000. *A.W.H. Phillips: Collected works in contemporary perspective*. Cambridge: Cambridge University Press.
Lucas, R.E. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit, and Banking* 12: 696–715.
Mäki, U. 1992. On the method of isolation ineconomics. In *Idealization IV: Intelligibility in science*, ed. C. Dilworth. Amsterdam: Rodopi.
Mäki, U. (ed.). 2002. *Fact and fiction in economics*. Cambridge: Cambridge University Press.
Mäki, M. (ed.). 2007. *The methodology of positive economics: Milton Friedman's essay fifty years later*. Cambridge: Cambridge University Press.
Marshall, A.W. 1890. *Principles of economics*, 8th ed. London: Macmillan. 1930.
McMullin, E. 1985. Galilean idealization. *Studies in History and Philosophy of Science* 16: 247–273.
Morgan, M.S. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.
Morgan, M.S. 2001. Models, stories and the economic world. *Journal of Economic Methodology* 8: 361–84. Repr. in Mäki (2002).
Morgan, M.S. 2002. Model experiments and models in experiments. In *Model-based reasoning: Science, technology, values*, ed. L. Magnani and N. Nersessian. New York: Kluwer Academic/Plenum Press.
Morgan, M.S. 2004a. Imagination and imaging in economic model-building. *Philosophy of Science* 71: 753–766.
Morgan, M.S. 2004b. Simulation: The birth of a technology to create 'evidence' in economics. *Revue d'Histoire des Sciences* 57: 341–377.
Morgan, M.S. 2006. Economic man as model man: Ideal types, idealization and caricatures. *Journal of the History of Economic Thought* 28: 1–27.
Morgan, M.S. 2007. The curious case of the Prisoner's Dilemma: Model situation? Exemplary narrative? In *Science without laws: Model systems, cases, and exemplary narratives*, ed. A. Creager, E. Lunbeck, and N. Wise. Durham: Duke University Press.
Morgan, M.S. 2008. The world in the model.
Morgan, M.S., and M. Morrison. 1999. *Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.
Morrison, M., and M.S. Morgan. 1999. Models as mediating instruments. In Morgan and Morrison.
Nowak, L. 1994. The idealization methodology and econometrics. In Hamminga and De Marchi.
Orcutt, G.H. 1960. Simulation of economic systems. *American Economic Review* 50: 893–907.
Pigou, A.C. 1929. The function of economic analysis. The Sidney Ball Lecture, University of Oxford, May. In *Economic essays and addresses*, ed. D.H. Robertson. London: P.S. King. 1931.
Qin, D. 1993. *The formation of econometrics*. Oxford: Clarendon Press.
Rappaport, S. 1998. *Models and reality in economics*. Cheltenham: Edward Elgar.
Robbins, L. 1932. *An essay on the nature and significance of economic science*. London: Macmillan.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *Review of Economics and Statistics* 21: 75–78.
Shubik, M. 1959. *Strategy and market structure*. New York: Wiley.

Shubik, M. 1960a. Bibliography on simulation, gaming, artificial intelligence and allied topics. *Journal of the American Statistical Association* 55: 736–751.
Shubik, M. 1960b. Simulation of the industry and the firm. *American Economic Review* 50: 908–919.
Slutsky, E.E. 1927. The summation of random causes as the source of cycle processes. *Econometrica* 5(1937): 105–146.
Sugden, R. 2000. Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology* 7: 1–31.
Sutton, J. 2000. *Marshall's tendencies: What can economists know?* Cambridge, MA: MIT Press.
Tinbergen, J. 1937. *An econometric approach to business cycle problems*. Paris: Hermann.
Tinbergen, J. 1939. *Statistical testing of business cycle theories*. Geneva: League of Nations.
Von Thünen, J.H. 1826. *Der Isolierte Staat.* Hamburg: Perthes. Trans. C. Wartenberg as *Von Thünen's isolated state.* Oxford: Pergamon, 1966.
Weber, M. 1904. 'Objectivity' in social science and social policy. In *The methodology of the social sciences.* Trans. and ed. E.A. Shils and H.A. Finch. New York: Free Press, 1949.
Weber, M. 1913. *The theory of social and economic organisations*, Trans. A.M. Henderson and T. Parsons as Part I of *Wirtschaft und Gesellschaft.* New York: Free Press, 1947.
Weintraub, E.R. 2002. *How economics became a mathematical science*. Durham: Duke University Press.

Models and Theory

Vivian Walsh

Ernest Nagel once remarked that '[t]he only point that can be affirmed with confidence is that a model for a theory is not the theory itself' (Nagel 1961, p. 116). And R.B. Braithwaite warned against the danger that: 'The theory will be identified with a model for it . . .' (Braithwaite 1953, p. 90). It will be argued here that Nagel, Braithwaite and the school of which they were representative were right to insist on a model/theory distinction, but wrong as to the nature of that distinction and the reasons for adopting it. The now defunct school referred to was christened by Hilary Putnam the 'Received View' (Putnam 1962). The Received View in the philosophy of science was (roughly) the logical positivist interpretation of science. It involved a model/theory distinction in an essential way. Logical positivist ideas penetrated economic theory and lived on there long after the fall of the Received View. The latter, after more than 30 years of dominance, came under such severe attacks that by the end of the 1960s, as Frederick Suppe later remarked, these attacks 'had been so successful that most philosophers of science had repudiated the Received View' (Suppe 1977, p. 618).

Now insofar as any economic theorists still harbour logical positivist ideas, they are committed (if consistent) to support a model/theory distinction. But this support, it is claimed, is given for the wrong reasons. To see this, it is necessary to distinguish the kind of model/theory distinction characteristic of the Received View from that advocated here. The role of models in the Received View was clearly put by Nagel. He distinguishes three components of a theory:

> (1) an abstract calculus that is the logical skeleton of the explanatory system . . .; (2) a set of [correspondence] rules that in effect assign an empirical content to the abstract calculus by relating it to the concrete materials of observation and experimentation; and (3) an interpretation or model for the abstract calculus, which supplies some flesh for the skeletal structure in terms of more or less familiar conceptual or visualizable materials (Nagel 1961, p. 90).

Crucial to this view is the concept of an 'abstract calculus', regarded as an axiomatized system of uninterpreted sentences. Carl G. Hempel, once a leading protagonist of the Received View, who has 'come to feel increasing doubts about its adequacy . . .' (Hempel 1977, p. 247), expresses his rejection of the idea thus:

> The conception of an uninterpreted calculus C seems to me misleading because it suggests that the basic assumptions of a theory . . . are expressed exclusively by means of the 'new' theoretical terms introduced by the theory . . . . Actually, however, the internal principles of most theories characterize the theoretical scenario at least in part by means of terms taken from the antecedent vocabulary (Hempel 1977, p. 250).

Consider how the Received View would have interpreted Gerard Debreu's canonical work,

Theory of Value (1959). At the end of chapter 2, for example, Debreu sums up the concepts developed 'in the language of the theory . . .' (p. 35); '[a]ll that precedes this statement is irrelevant for the logical development of the theory. Its aim is to provide possible interpretations of the latter'. The summary would thus have to be regarded (on the Received View) as part of an uninterpreted calculus. But it begins: '*The number L of commodities is a given positive integer. An action a of an agent is a point of R_L, the commodity space. A price system* . . . (Debreu 1959, p. 35; emphasis in original). Here we have the familiar terms commodity, action, commodity space, and price system. Hempel, considering terms like mass, energy and momentum, remarks: 'It might be replied that when antecedently available terms are thus used in the formulation of a theory they function in quite novel principles and accordingly acquire totally new meanings, and that they should therefore be reckoned among the theoretical terms' (Hempel 1977, p. 250). He demolishes this claim for the case of physics. Readers may likewise consider whether commodity, action, price etc., are wholly uninterpreted concepts in Debreu's formalization. Yet Debreu's book is surely a canonical example of *formalization* in economic theory.

The treatment of a model as an 'interpretation' of a calculus arose because of epistemological views which show clearly in the stipulation that a set of 'correspondence rules' and an 'uninterpreted' calculus be considered part of the theory. Typical of the Received View, these need not concern us now. The model/theory distinction which can be useful in economics does not turn on epistemological chastity about some pure 'uninterpreted' calculus. This can be illustrated informally by considering the case of general equilibrium.

If the models for a theory are isomorphic in their structure (the criteria for this depending on the subject matter), then the theory comes as close as possible to having only one model, and the theory is said to be *categorical.* Now the theory of general equilibrium is anything but categorical. Consider a strictly *minimal* set of criteria for calling a given mathematical structure 'a model for the theory of general equilibrium'. One might say M is a model for the theory of general equilibrium only if: (1) there is a non-empty set of agents, each endowed with goods and/or factor services in a non-negative quantity, and with preferences such that an undominated attainable option is chosen; (2) there is a set of quantity relations; (3) there is a set of price relations; (4) there is a set of duality conditions under which the system has at least one equilibrium. Note that terms like agent, goods, price etc. have their ordinary sense and are not *uninterpreted* in the sense of the Received View. Models of general equilibrium are then different determinations of the possibilities left open in the above definition. But this is a purely workaday filling in of one or other detailed determination to an already (epistemologically) *interpreted* framework.

The interest lies in the fact that these additions can lead to Strikingly different models, and many of them. A general equilibrium model may encompass only pure exchange, yet even within this class many strikingly different models exist, some with an unchanging competitive core, some with disequilibrium trading, some with quantity rationing, some with money. If a model encompasses production, this may involve only the services of given resources, or the reproduction of commodities by means of the same commodities, or both produced and non-produced inputs. The class of commodity reproduction models, again, includes models in an equilibrium defined in terms of the duality condition of a uniform rate of profit. But there are also models which are in quantity rationed quasi equilibria, in a process of gravitation to reproduction prices (Duménil and Lévy 1984). And there are models of continued reproduction which are in dynamic structural change, with equilibrium conditions allowing *different* rates of profit (Pasinetti 1981).

Certain themes may play a major part in some models and a minor part, or none, in others. Demand plays a major role in models of pure exchange, models with production from given resources, certain gravitation models, and in structural dynamics. It plays a minor role in commodity reproduction models in stationary or steadily growing states. Again, in some models all agents treat prices as given, in others (such as some quantity constrained exchange models and some gravitation models) certain agents set prices.

A pair of models may clearly have properties which make it impossible to combine them, or to reduce one to the other, yet both may be models for the theory of general equilibrium. Yet sometimes a particular class of models is identified with the theory. Thus the neo-Walrasian class of models is sometimes treated as the theory of general equilibrium. (This might have been forgivable in the 1920s, when no other important models were widely known.) A reductionist argument is then attempted, aimed at showing that some other models are simply mis-specified neo-Walrasian models.

Where a new class of model is not for some reason found objectionable, however, the reductionist move is seldom attempted. Consider the fast growing class of models with non-market clearing prices, conjectures and the like, which have been significantly called by Frank Hahn 'non-Walrasian equilibria' (Hahn 1978). Hahn does not seem to want to prove that these are just a special case of Walrasian equilibria. On the other hand he has made herculean efforts to show that what he calls 'Neo-Ricardianism' is contained in 'neoclassical economics' (Hahn 1982, pp. 353–74), which is identified in effect with the neo-Walrasian class of models for general equilibrium theory.

As has been remarked by Frederick Suppe, '[t]his view of scientific development, which I call the *thesis of development by reduction,* and the Received View clearly go hand in hand' (Suppe 1977, p. 56, emphasis in original). May we hope that this kind of reductionist argument will go hand-in-hand with the Received View into intellectual history?

See Also

- ▶ Axiomatic Theories
- ▶ Methodology
- ▶ Philosophy and Economics

Bibliography

Braithwaite, R.B. 1953. *Scientific Explanations: A study of the function of theory, probability and law in science*. Cambridge: Cambridge University Press.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*, Cowles Foundation monograph No. 17. New York: Wiley.

Duménil, G., and D. Lévy. 1984. Une restauration de l'analyse classique de la dynamique concurrentielle. In *La Gravitation*, Systèmes de Prix de Production, vol. 2, 3, ed. C. Bidard and R.C.P. Cahiers de la. Nanterre: University of Paris.

Hahn, F.H. 1978. On non-Walrasian equilibria. *Revue of Economic Studies* 45(1): 1–17.

Hahn, F.H. 1982. The neo-Ricardians. *Cambridge Journal of Economics* 6(4): 353–374.

Hempel, C.G. 1977. Formulation and formalization of scientific theories, a summary abstract. In Suppe (1977).

Nagel, E. 1961. *Structure of science: Problems in the logic of scientific explanation*. New York: Harcourt Brace.

Pasinetti, L.L. 1981. *Structural change and economic growth*. Cambridge: Cambridge University Press.

Putnam, H. 1962. What theories are not. 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.

Suppe, F. (ed.). 1977. *The structure of scientific theories*, 2nd ed. Urbana: University of Illinois Press.

Models of Growth

H. Uzawa

The Harrod–Domar Model

The year 1939 was marked by the appearance of Harrod (1939) which gave a major impetus to the development of growth theory. Harrod was concerned with the problem of probable inconsistency between the conditions of full employment and a steady state of economic growth. The conditions under which full employment is secured are necessarily of a short-run nature, while a steady state of growth requires certain fundamental dynamic equations to be satisfied.

One of the fundamental equations introduced by Harrod expresses the equilibrium of a steady state of growth:

$$g_w c_r = s,$$

where g_w is the warranted rate of growth, c_r is the required capital coefficient, and s is the saving coefficient. Harrod's warranted rate of growth g_w is defined as

> the rate of growth, if it occurs, will have satisfied all members of the economy, while the required capital coefficient c_r is defined as the requirement for new capital divided by the increment of total output to sustain which the new capital is required.

Harrod assumed that the saving ratio s is a constant, to be dependent upon the psychological and social characteristics of the economy. Under the assumptions of the neutrality of inventions and of the constancy of the rate of interest, the required capital coefficient c_r is also a constant. If the rate of growth g is higher than the warranted rate of growth g_w, then the capital coefficient c is lower than the required capital coefficient c_r. The accumulation of capital then would be insufficient to sustain a steady state of growth. On the other hand, if g is lower than g_w, c is higher than c_r. Some portion of capital would be necessarily left unutilised at a steady state of growth. Harrod thus gave a simple proof for the instability of processes of economic growth in a capitalist economy.

The analysis of the instability of the process of economic growth in a capitalist economy, as discussed by Harrod, was one of the major attempts to extend Keynes's *General Theory* and regarded as one of basic pillars upon which the modern growth theory has been built.

In the *General Theory,* Keynes attempted to formulate institutional arrangements of the modern capitalist economy in terms of a coherent macroeconomic analytical framework and showed that the allocative mechanism in a decentralized, private-enterprise market economy resulted in a state of involuntary unemployment, unless stabilizing fiscal and monetary policies are effectively utilised. Harrod's dynamic analysis may be regarded as an extension of the Keynesian analysis to cover the economy at a steady state of growth. However, it was after the end of World War II that the Keynes–Harrod analysis received fuller attention. Indeed, the problems of economic growth and full employment were at the centre of attention of the political and social planners in major capitalist countries, both developed and less developed, and the period of approximately twenty-five years up to the end of the 1960s may be regarded as one of stable economic growth, largely due to the adaption of what may be properly termed a Keynesian policy.

Harrod's analysis was further elaborated by Evsey Domar (1946), where some of the underlying assumptions in the Harrod's model were more explicitly brought out and the long-run implications were discussed in more detail. While Keynes was primarily concerned with the role of investment as an instrument for generating income, both Harrod and Domar focused their attention upon the effect of investment to increase productive capacity.

(a) The amount of capital and labour required to produce a unit of output are both technologically given.
(b) A constant fraction of income is saved.
(c) The rate of increase in labour forces is exogenously given.
(d) Inventions are neutral in the sense of Harrod and the rate of increase in labour efficiency is exogenously given.

Under these assumptions, the Harrod knife-edge instability of a steady state of economic growth was rigorously proved by Domar (1946). The equality of the natural rate of growth with the warranted rate of growth, on which the existence of a steady state of growth crucially hinges, occurs only for an economy for which the saving ratio, the capital coefficient, and the rate of increase in labour forces satisfy particular relationships. Any path of capital accumulation in such an economy generally exhibits an unstable feature; either involuntary unemployment tends to be increased without limit or capital continues to be accumulated in such a manner that the stock of unutilized capital piles up indefinitely.

Neoclassical Growth Models

The instability property of the process of capital accumulation inherent in the Harrod–Domar model may crucially hinge upon the nature of

the basic assumptions concerning technical and social structure of the capitalist economy in question.

In particular, the assumption of a constant capital coefficient seems to be a pivotal one in the Harrod–Domar analysis. The neoclassical growth models, developed by Tobin (1955), Solow (1956), Swan (1956), Ara (1958) and Meade (1961), among others, take an explicit note of the possibility of the substitution between capital and labour and conclude that growth paths in a capitalist economy have a trend to converge to a steady state.

A neoclassical growth model typically is formulated in terms of a one-commodity economy, where output is produced by two factors of production, capital and labour. Total output Y is given by the aggregate production function

$$Y = F(K, L)$$

where K and L represent inputs of capital and labour, respectively.

The possibility of substitution between capital and labour then is expressed by the assumption that the aggregate production function $F(\cdot,\cdot)$ is continuously differentiable so that the marginal rate of substitution between capital and labour is well defined. Constant returns to scale is also assumed so that the aggregate production function $F(\cdot,\cdot)$ is linear and homogeneous.

At each time t, real output $Y(t)$ is produced by using the stock of capital $K(t)$ and labour services $L(t)$. A constant portion of real income $Y(t)$ is assumed to be consumed and the rest is saved. Net investment is assumed to be equal to savings. If we denote by s the saving ratio, then the rate of accumulation of capital is given by

$$\frac{dK(t)}{dt} = sY(t) = sF[K(t), L(t)],$$

while the available labour is assumed to grow at a constant rate n:

$$\frac{dL(t)}{dt} = nL(t).$$

Growth paths in a neoclassical model then are completely described by these two differential equations, which may be, due to the constant returns to scale assumption, reduced to the following:

$$\frac{dk(t)}{dt} = sf[k(t)] - nk(t), \tag{1}$$

where $k(t) = K(t)/L(t)$ is the capital–labour ratio at time t and $f[k(t)] = F[k(t), l]$ is real output per capita at time t.

The assumption of diminishing marginal rates of substitution between capital and labour may be expressed by the concavity of the per capital output function $f(k)$; namely,

$$f''(k) < 0, \quad \text{for all} \quad k > 0.$$

Hence, the solution paths to the differential equation (1) tend to converge to the stationary state k^* of the system (1):

$$sf(k*) = nk^*.$$

The existence of the stationary state k^* is generally guaranteed, particularly when $f'(0) = \infty$ and $f'(\infty) = 0$.

The neoclassical growth models have many variants, in particular concerning the saving ratio assumption. While the constancy of the saving ratio s has been adapted in most of the neoclassical models, some have taken an explicit cognizance of the fact that it may depend upon the level of per capita real income and the rate of interest. The assumption that the rate of population growth is exogenously given has been cricially examined, particularly by Buttrick (1960). The stability property, however, has been verified in most of the neoclassical models.

Kaldor's 'Stylized' Facts

The perspective of the growth theory may be best illustrated by the six 'styled' facts put forward by Nicholas Kaldor (1961), which have been obtained by observing the process of economic growth in capital economies. They are (1) the continued

growth in the aggregate output and in the per capita output at an ever-increasing rate; (2) the capital-labour ratio has continuously increased; (3) the rate of profit on capital has been steady, significantly higher than the real rate of interest, at least for most of the more advanced capitalist economies; (4) the steady capital coefficient has been maintained; (5) the share of investment in output has been highly correlated with the share of profits in income; (6) the divergence of the long-run rate of increase in labour productivity and of the aggregate output in different economies.

Some of Kaldor's 'stylized' facts may not be necessarily borne out by the observed statistical data, particularly in the latter half of the 20th century. However, they may be taken as a convenient starting point for the construction of theoretical growth models. In the light of Kaldor's 'stylized' facts, both the Harrod–Domar model and neoclassical growth models may need a re-examination of the basic assumptions. Particular attention was paid to the apparent inconsistency between the continued increase in the capital–labour ratio and a constant capital-coefficient. This indeed was one of the problems Harrod addressed himself in Harrod (1937), and later elaborated in Joan Robinson (1937–8). It is related to the role which inventions have played in the process of economic growth. A technical invention was defined by Harrod to be neutral if the optimum capital coefficient remains constant when the rate of interest is kept constant. It was shown by Joan Robinson that if a technical invention is neutral in Harrod's sense, then the increase in the efficiency of labour is determined independently of the stock of capital being utilized. In terms of the aggregate production function, the characterization of the Harrod neutrality was explicitly brought out in Uzawa (1961a). Let technological conditions change over time, so that the aggregate production function may be represented by

$$Y = F(K, L, t).$$

Then it was proved that technical inventions are neutral in the sense of Harrod if and only if the aggregate production function $Y = F(K, L, t)$ is written as

$$Y = G[K, A(t)L],$$

where $A(t)$ indicates the efficiency measure for labour at time t, to be determined independently of K and L.

It was then shown in Uzawa (1961a) that, if technical inventions are neutral in the sense of Harrod, the steady state of the neoclassical growth model is characterized by the conditions that the capital coefficient remains constant while the capital–labour ratio continues to increase at the rate equal to that of labour efficiency, and paths of economic growth necessarily converge to the steady state.

In the neoclassical growth models, the rate of profit has been largely identified with the rate of interest. At the same time, savings were regarded as determined by total income, largely independently of the way total income is divided between the factors of production. These properties are related to the way the working of an economic system is viewed. In the traditional neoclassical economic theory, the working of economic activities is described by the representative *homo economicus* who behaves himself in accordance with the subjective value judgement he possesses independently of the economic environments and historical and social circumstances. The representative *homo economicus* acts as a producer and a consumer at the same time, and he is the owner of all the factors of production, including labour. He divides his income between consumption and savings in such a manner that his intertemporal preference ordering is satisfied. The aggregate savings then are channelled into investment. Full employment necessarily results in such a neoclassical economy, and investment is automatically determined by the amount of savings.

Marxian and Kaldorian Growth Models

Unlike the neoclassical growth models, the Marxian and Keynesian growth models have been built upon the basic premises that a capitalist

economy is composed of different, occasionally conflicting, classes, and patterns of economic growth would reflect the interaction of classes in the process of resource allocation and income distribution. A typical Marxian growth model is the one where the neoclassical production function is assumed to summarize the production processes, but the amount of savings depends upon the way total product is divided between wages and profits. The accumulation of capital then is given by dK dt

$$\frac{\mathrm{d}K}{\mathrm{d}t} = s_P P + s_W W,$$

where P and W stand for profits and wages, respectively, and s_P and s_W are the average propensities to save out of profits and wages, respectively.

The simplest Marxian case may be represented by the conditions: $s_P = 1$ and $s_W = 0$. The stability of growth paths has been shown for the general case when

$$0 \leqq s_w < s_p \leqq 1,$$

and profits P and wages W are determined by marginal products capital and labour, respectively.

Kaldor (1956) introduced a slightly different model, in which the distribution of income Y between profits P and wages W is so determined as to equate the forthcoming savings S with investment I, the latter being independently determined by entrepreneurs. Namely, profits P and wages W are determined by the following two equations:

$$\begin{aligned} Y &= \mathrm{P} + W \\ I &= s_P P + s_W W, \end{aligned}$$

where Y and I are exogenously given.

The stability of growth processes in a model where distribution is determined by what Kaldor has termed the Keynesian theory of distribution is related to the way entrepreneurs decide total investment I.

The Marx–Kaldor theory of economic growth was further elaborated by Pasinetti (1962).

Two-Sector Growth Models

The neoclassical growth models have been based upon the concept of the aggregate production function which relates the total output measured in terms of a certain homogeneous quantity to the inputs of labour and capital. A number of attempts were made to extend the analysis to cover the situation where there exist various types of goods which are produced by different technologies. Particular attention was paid to the two-sector growth models, where there are two types of goods, investment goods and consumption goods, to be produced by two factors of production, labour and capital.

The simple case where both goods are produced with constant coefficient technologies was discussed by Shinkai (1960). Shinkai's main conclusion was to relate the stability of the growth process in such a two-sector model to the relative intensities of two goods. Shinkai's model was extended to the case in which substitution between capital and labour is possible in the production of both investment goods and consumption goods (Meade 1961; Uzawa 1961b, 1963).

The basic premises upon which the two-sector growth models have been built may be briefly summarized. Two-sector growth models consider an economy in which there are two sectors, one producing consumption goods and the other investment goods, to be labelled C and I respectively. Both consumption goods and investment goods are composed of homogeneous quantities and produced by two homogeneous factors of production, labour and capital. Consumption goods are instantaneously consumed, while capital goods, being the accumulation of investment goods, depreciate at a fixed rate. In each sector, production is subject to constant returns to scale and diminishing marginal rates of substitution between capital and labour. Joint products are excluded and neither external economies nor dis-economies exist. The quantity of each good to be produced is related to the quantities of capital and labour to be allocated in each sector. The production function in each sector then is assumed to be a linear homogeneous, continuously differentiable function of two variables, capital and labour.

At each moment, total quantity of capital available to the economy is determined as the result of past investment, while the available labour forces are assumed to be exogenously given, to grow at a certain fixed rate. Then the quantities of capital and labour allocated to the two sectors are constrained by the quantities of capital and labour available at that moment.

In two sector growth models, both outputs and factors of production are allocated in perfectly competitive markets, so that in each sector the wage is equal to the marginal product of labour and the rentals of capital to the marginal product of capital. In each sector, the optimum capital–labour ratio then is uniquely determined by the wage–rental ratio. Consumption goods are defined to be always relatively more (or less) capital intensive than investment goods if the optimum capital–labour ratio is higher (or lower) in the *C*-sector than in the *I*-sector for all possible wage–rentals ratios.

The allocation of capital and labour between two sectors is uniquely determined if the relative price of two goods is given. If consumption goods are more capital intensive than investment goods, then the relative price of consumption goods in terms of investment goods will be increased when the wage–rentals ratio is decreased.

The relative price of two goods will be determined once the demand conditions are specified.

In a Marxian situation where labourers consume all their wages and capitalists save all their profits, the short-run equilibrium will be uniquely determined if consumption goods are always more capital intensive than investment goods. Paths of growth equilibrium have been shown to be stable under the same capital-intensity condition.

On the other hand, in a neoclassical economy where a fixed proportion of total income is spent on consumption and the rest is saved, the short-run equilibrium has been shown to be uniquely determined, regardless of the capital-intensity condition. However, the stability of growth of growth equilibrium is established only for the case where consumption goods are always more capital intensive than investment goods.

The concept of capital utilized in the two-sector growth models has been based upon the neoclassical theory in the sense its use can be shifted from one sector to another without incurring any additional cost or any time lag. The model developed in Inada (1966) recognizes that most of capital embodied in modern technologies cannot be freely shifted from one sector to another and has to stay in the sector where it has been invested.

This is related to the line of model construction, which may be traced back to Fel'dman (1928), and paves a way to the development of the Keynesian theory of economic growth.

As for growth models with heterogeneous capital goods, a number of what may be termed vintage models have been constructed and their implications on the pattern of growth equilibrium have been analysed in detail. A particular interest is with the one built by Solow (1960). Solow's model considers an economy which is composed of homogeneous labour and capital equipment of various vintages, each of which embodies the technologies of the time when it is built. If the technical progress embodied in vintage capital in Solow's model is neutral in the sense of Harrod, then it has been proved that any path of growth equilibrium asymptotically approaches the steady state where the vintage distribution of capital equipment remains stationary (Uzawa 1964b).

Further contributions to the theory of vintage capital and the related topic of induced investment have been made by Atkinson and Stiglitz (1969), Phelps (1966), Kennedy (1964), and Leif Johansen (1959).

Optimum Economic Growth

One of the basic problems in economic planning, in particular in underdeveloped countries, is concerned with the rate at which society should save out of current income to achieve an optimum growth. It is closely related to the problem of how to allocate scarce resources at each moment of time between the production of consumption goods and investment goods. It was analysed within the context of the two-sector growth models, as introduced in Meade (1961), Srinivasan (1964), and Uzawa (1964a). The Srinivasan–Uzawa analysis focused its attention on evaluating the impact of roundabout methods of production upon the welfare of the

society, as expressed by a discounted sum of per capita consumption over time. It abstracted from the complications that would arise by taking into account those factors such as changing technology and structure of demand, the role of foreign trade (in particular of capital movements) and tax policy that may be generally regarded as decisive in the course of economic development. It is postulated that a certain quantity of consumption goods per capita is required to sustain a given rate of population growth. The constraint will become effective for an economy with relative shortage of capital, and it results in the phenomenon of 'the vicious circle of poverty'.

The structure of optimum paths of capital accumulation differs significantly according to whether consumption goods are relatively more capital-intensive or less capital-intensive than investment goods. If consumption goods are always more capital-intensive than investment goods then there exist two critical–labour ratios, k_I^* and k_C^* such that, if the initial capital–labour ratio of the economy is less than k_I^*, then, along the optimum path, the economy produces just enough consumption goods to meet the minimum requirements and devotes the rest of scarce resources in the production of investment goods until the time when the economy's capital–labour ratio reaches the capital ratio k_I^*, and from then on it proceeds to produce both investment goods and consumption goods, keeping the imputed price of the two outputs constant and approaching the stationary state. On the other hand, if the initial capital–labour ratio of the economy is larger than the critical ratio k_C^*, then, along the optimum path, the economy is specialized to the production of consumption goods until the capital–labour ratio is reduced to the critical ratio k_C^*. When the critical ratio k_C^* is reached, then both consumption goods and investment goods are produced, asymptotically approaching the stationary state.

Two-Class Models of Economic Growth

Most of the growth models described above have been built upon premises directly involving aggregate variables, without specifying the postulates which govern the behaviour of individual units comprising the national economy. In particular, the specifications of aggregate savings have seldom been based upon analysis of rational behaviour concerning savings and consumption. Similarly, the aggregate behaviour of investment has not been derived from the rational behaviour of business firms; instead, it has been postulated in terms of *ad hoc* relations involving market rate of interest, rate of profit, and other variables. In Uzawa (1969), an attempt was made to build a formal model of economic growth for which the aggregate variables such as consumption, savings and investment are described in terms of individual units' rational behaviour. A private-enterprise economy is divided into two sectors; the household sector and the corporate sector. Households decide how to consume goods and services produced in the corporate sector; they are endowed with labour and possess the securities issued by the corporate sector. A business firm in the corporate sector consists of a complex of fixed factors of production, such as factories, machinery, and others, including managerial abilities and technological skills. Real capital is regarded as an index to measure the productive capacity of such a complex of capital goods endowed within the firm at each moment of time; it is increased as the stock of fixed factors of production is accumulated as the result of investment activities. The relationships between real investment and the resulting increase in real capital may be characterised by what may be called the Penrose curve, which incorporates the basic tenure of the analysis expounded by Edith Penrose (1959). Each business firm plans the levels of employment and investment in such a way that the discounted present value of expected future net cash flows is maximized. The desired level of investment then depends upon the expected rate of profit and the market rate of interest.

The behaviour of an individual household may be analysed in terms of Irving Fisher's theory of time preference, as formulated by Koopmans (1960). The marginal rate of substitution between current and future consumption is represented by the Fisherian schedule, which relates the rate of time preference to the current level of consumption and to the utility level for all future

consumption. The optimum propensities to consume and save are then derived as functions of the expected market rate of interest and permanent income.

These analyses are put together to formulate a two-class model of economic growth, where the stability of the short-run and long-run equilibrium is discussed in terms of the Penrose curve and the Fisherian schedule.

Keynesian Models of Economic Growth

Whether or not the dynamic allocation of scarce resources through the market mechanism may achieve stable economic growth is not simply a matter of theoretical interest, but is indispensable in the discussion of the effect of public policy. There are two opposing approaches to the problem of dynamic stability of the market mechanism. One approach is based upon the analytical framework of neoclassical economic theory, and the other is discussed in terms of the Keynes–Harrod theory of economic dynamics. The neoclassical approach derives the conclusion that the process of equilibrium growth in a market economy is dynamically stable and the conditions of full employment generally prevail. The Keynes–Harrod approach, on the other hand, concludes that the market allocation of scarce resources is inherently instable in a modern capitalist economy and that maintaining stable economic growth, together with full employment and price stability, is akin to walking on the edge of a knife. It may be worthwhile to examine the reason why these two opposing conclusions concerning the stability of the growth process in a market economy are obtained.

One of the crucial elements which distinguish one approach from the other is concerned with the concept of capital. In the neoclassical approach, capital refers to the various factors of production which have been accumulated through refraining from consumption in the past. Production is carried out by utilizing capital together with labour and other variable factors of production which are obtained through markets. In the neoclassical theory, the phenomenon of the fixity of capital has not been handled explicitly, so that the market price of the stock of capital is the same for newly produced capital goods and for existing capital goods which are the result of investment in the past. Any member of the economy may engage in productive activity either by purchasing capital goods or by renting the services of capital goods, and at the same time may engage in consumption activity, resulting in the disappearance of the essential difference between consumers and producers. Accordingly, various members of the economy may hold either physical or financial assets in whatever manner they prefer. Investment, as an accumulation of fixed capital, loses its essential meaning, and the difference between rate of interest and rate of profit disappears. Markets for outputs and factors of production are assumed to be perfectly competitive.

Under these neoclassical assumptions, Say's Law is shown to hold true and full employment necessarily results. Growth equilibrium is generally shown to be dynamically stable.

The stability of monetary growth in neoclassical theory has been similarly handled, particularly by Tobin (1965), Sidrauski (1967), Harry Johnson (1966), Levhari and Patinkin (1968) and Uzawa (1974). The neoclassical theory of monetary growth typically ignores the institutional details of the mechanism by which money is supplied by the central bank, and money is assumed to be distributed to the economic units of the economy through transfer payments. In neoclassical theory, money performs the function of a consumer good, contributing to the increase in the level of utility for each individual, and it may also serve the role of a factor of production, increasing the marginal products of real factors of production. The demand for money thus may be assumed to depend upon the market rate of interest and the level of income.

The aggregate demands for real capital and money are related to the price level, and the equilibrium price level then is determined as the level at which the demand for the holding of money balances is equated to the supply of money. The rate of interest then becomes the real rate of interest plus the expected rate of price increase. The stability of monetary growth in the neoclassical setting is thus related to the way expectations concerning future price changes are formed. If

expectations are adjusted according to adaptive expectations of the Cagan–Nerlove type, then equilibrium growth in the neoclassical model of monetary growth is dynamically stable provided the speed of adjustment in expectations is relatively small, as one would expect from similar analyses such as Cagan (1956).

The Keynesian model of monetary growth, on the other hand, may be formulated in terms of the two-class economy as briefly outlined above. The private sector consists of households and business firms, and the government sector provides various public goods and services which are financed through taxes or the issue of money. Money supply is increased to meet fiscal deficits, but money supplied through open market operations changes the pattern of portfolio balances in the economy.

The market system is divided into three types; the goods and services market, the labour market, and the financial market. The goods and services market and the financial market both are assumed to be instantaneously adjusted to the equilibrium positions. In the labour market, however, when the demand for labour exceeds the supply, the money wage rate is instantaneously adjusted to the equilibrium level, but when the demand for labour is less than the supply, the money wage rate remains at the current level, resulting in involuntary unemployment. Money and short-term securities are dealt with in efficiently organized markets, but price adjustments for long-term securities are not necessarily efficient and there is a time lag in the adjustment of long-term securities prices.

The schedule of the aggregate supply price is then defined; it relates the aggregate amount of goods and services measured in wage-units to total employment in such a manner that entrepreneurs' profits are maximized subject to the constraints imposed by the conditions prevailing in the economy. On the other hand, aggregate demand is determined by the behaviour of households, business firms and government concerning consumption and investment.

Equilibrium in the goods and services market is obtained when aggregate supply is equal to aggregate demand. The level of total employment at the equilibrium does not correspond to the full employment level, resulting in the situation of involuntary unemployment. The effective demand is closely related to the level of investment, which in turn is influenced by the market rate of interest in the long-term securities market.

The market rate of interest is determined by the equilibrium conditions in the markets where money and short-term securities are transacted. Thus, in order for the economy to sustain the conditions of full employment and continuous economic growth, certain conditions have to be satisfied between the long-term rate of interest, the rate of increase in money supply, and the rate of increase in productive capacity of the economy. It is then possible to prove that equilibrium growth paths in such a Keynesian model tend to exhibit an instability of the Harrod knife-edge type; along any growth path, either the level of employment remains at the level below full employment or the price level tends to increase with an accelerating rate. To stabilize the process of economic growth, it becomes necessary to adopt a flexible policy concerning the supply of money which is directed toward stabilization of the market rate of interest or the rate of increase in the price level.

The phenomenon of economic growth exhibits a quite different picture, according to whether we take the neoclassical approach or the Keynesian approach. In the neoclassical growth model, the path of economic growth is dynamically stable under fairly general conditions, while in the Keynesian model there is an intrinsic tendency for the process of growth to be dynamically unstable unless stabilizing monetary and fiscal policies are adopted.

See Also

- ▶ Harrod–Domar Growth Model
- ▶ Multisector Growth Models
- ▶ Ramsey Model
- ▶ Turnpike Theory

Bibliography

Ara, K. 1958. Capital theory and economic growth. *Economic Journal* 68: 511–527.

Arrow, K.J. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29: 155–173.
Atkinson, A., and J.E. Stiglitz. 1969. A new view of technical change. *Economic Journal* 79: 573–578.
Buttrick, J.A. 1960. A note on growth theory. *Economic Development and Cultural Change* 9: 75–82.
Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.
Domar, E. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14: 137–147.
Harrod, R.F. 1937. Review of Joan Robinson's *essays in the theory of employment. Economic Journal* 47: 326–330.
Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.
Inada, K. 1966. Investment in fixed capital and the stability of growth equilibrium. *Review of Economic Studies* 33: 19–30.
Johnson, H.G. 1966. The neo-classical one-sector growth model: A geometric exposition and extension to a monetary economy. *Econometrica* 33: 265–287.
Jorgenson, D.W. 1961. Stability of a dynamic input–output system. *Review of Economic Studies* 28: 105–116.
Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23(2): 83–100.
Kaldor, N. 1961. Capital accumulation and economic growth. In *The theory of capital*, ed. F.A. Lutz and D.C. Hague. London: Macmillan.
Kennedy, C. 1964. Induced bias in innovation and the theory of distribution. *Economic Journal* 74: 541–547.
Koopmans, T.C. 1960. Stationary ordinal utility and impatience. *Econometrica* 28: 287–309.
Koopmans, T.C. 1964. On a concept of optimum economic growth. *Pontificiae Academiae Scientiarium Scripta Varia* 28: 225–287.
Levhari, D., and D. Patinkin. 1968. The role of money in a simple growth model. *American Economic Review* 58: 713–753.
Meade, J.E. 1961. *A neoclassical theory of economic growth*. New York: Oxford University Press.
Morishima, M. 1964. *Equilibrium, stability and growth: A multi-sectoral analysis*. Oxford: Oxford University Press.
Pasinetti, L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29: 267–279.
Penrose, E.T. 1959. *The theory of the growth of the firm*. Oxford: Blackwell.
Phelps, E.S. 1966. Models of technical progress and the golden rule of research. *Review of Economic Studies* 33: 133–145.
Ramsey, F.P. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Robinson, J. 1938. The classification of inventions. *Review of Economic Studies* 5: 139–142.
Shinkai, Y. 1960. On the equilibrium growth of capital and labor. *International Economic Review* 1: 107–111.
Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.
Solow, R.M. 1960. Investment and technical progress. In *Mathematical methods in the social sciences 1959*, ed. K.J. Arrow and S. Karlin. Stanford: Stanford University Press.
Srinivasan, T.N. 1964. Optimal savings in a two-sector model of economic growth. *Econometrica* 32: 358–373; errata 33, April, 474.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Tobin, J. 1955. A dynamic aggregative model. *Journal of Political Economy* 63: 103–115.
Uzawa, H. 1961a. Neutral inventions and the stability of growth equilibrium. *Review of Economic Studies* 28: 117–124.
Uzawa, H. 1961b. On a two-sector model of economic growth. *I. Review of Economic Studies* 29: 40–47.
Uzawa, H. 1963. On a two-sector model of economic growth, II. *Review of Economic Studies* 30: 105–118.
Uzawa, H. 1964a. Optimal growth in a two-sector model of capital accumulation. *Review of Economic Studies* 31: 1–24.
Uzawa, H. 1964b. A note on Professor Solow's model of technical progress. *Economic Studies Quarterly* 15: 63–68.
Uzawa, H. 1969. Time preference and the Penrose effect in a two-class model of economic growth. *Journal of Political Economy* 77: 628–652.
Uzawa, H. 1974. On the dynamic stability of economic growth: the neoclassical versus Keynesian approaches. In *Trade, stability, and macroeconomics*, ed. P.A. Samuelson and L. Horwich. New York: Academic.

Modern Money Theory

L. Randall Wray

Abstract

Modern Money Theory (MMT) is a relatively new approach to macroeconomics that focuses on building an understanding of the operation of sovereign currency systems and on developing a policy framework based on that understanding. This article first summarises the main conclusions of MMT – the most important of which is that a nation that issues its own sovereign currency does not face financial or solvency constraints. We next trace the

intellectual antecedents of MMT, which rest 'on the shoulders of giants'. MMT revives the State Theory of Money (or Chartalism) and integrates it with a variety of heterodox approaches to macroeconomics, including the credit, circuitiste and endogenous approaches to money, the functional finance approach to budgeting, the financial instability hypothesis and the sectoral balances approach. Among those giants, MMT borrows from Knapp, Keynes, Innes, Schumpeter, Lerner, Minsky and Godley. This article shows how their theories have been integrated by MMT to provide a coherent approach to macroeconomic theory and policy. In the final section we summarise the main implications for policy-making.

Keywords

Abba Lerner; A. Mitchell Innes; Central bank independence; Circuit approach; Endogenous money; Eurozone; Exchange rate regimes; Financial instability; Fiscal policy; G. F. Knapp; Government deficits; Hyman Minsky; J. A. Schumpeter; Modern money theory; Monetary policy; Sovereign currency; State theory of money; Taxes drive money; Wynne Godley

JEL Classifications

B5; B22; B31; E12; E40; E58; E62

Introduction: The Basics of MMT

Modern Money Theory (MMT) is a relatively new approach to macroeconomics that focuses on building an understanding of the operation of sovereign currency systems and on developing a policy framework based on that understanding. The earliest expositions of MMT were Mosler (1995) and Wray (1998); a recent 'primer' on MMT is Wray (2012; new edition 2015). In this section we begin with a definition of sovereign currency and then summarise the basic conclusions of MMT.

Sovereign Currency

A sovereign currency system is one in which the government issues its own currency denominated in its money of account. Typically, it denominates taxes and other obligations in the same currency, and its courts enforce contracts in the official money of account. Private entities normally denominate most money contracts as well as prices in the sovereign's money of account. Across the world and throughout recorded history, most nations have adopted sovereign currency systems. Examples of sovereign currency systems include the USA, the UK, Mexico, Russia and South Africa.

A sovereign currency system can be contrasted with one in which the government adopts a foreign currency or operates with a currency board arrangement. For example, a number of countries today have adopted the US dollar for domestic use, such as Panama, Ecuador and El Salvador, or issue their own currency convertible to the US dollar. Others have operated currency boards based on foreign currencies, such as Hong Kong and Argentina in the 1990s (both pegged to the dollar, although Argentina abandoned the dollar in 2002 in the depths of a crisis), and Singapore (which uses an undisclosed basket of currencies). With a traditional currency board, the country promises to convert its currency to the foreign currency on demand at a fixed rate. By far the boldest experiment of operating without a sovereign currency system is the entire European Monetary Union, in which each member nation dropped its currency and adopted a 'foreign' currency, the euro. The classical gold standard – in which the currency is convertible on demand to bullion at a fixed exchange rate – is another example of a non-sovereign currency system.

Countries that operate with sovereign currency systems can choose to manage their exchange rates. At one end of the spectrum, a country can adopt a floating exchange rate. At the other end, it promises to convert its own currency to a foreign currency (or to precious metal) at a fixed exchange rate. Most countries operate somewhere between the two extremes, with no promise to convert but managing the exchange rate within an informal

range. Note that a country that operates with a firm peg (to gold or foreign currency) can be analysed as a non-sovereign currency system.

MMT argues that when a country issues its own currency and operates with a flexible exchange rate it obtains the greatest degree of domestic policy space. Because it does not promise to provide a foreign currency (or gold) in exchange for its currency, it has greater freedom to use monetary and fiscal policy to achieve domestic goals. If it pegs, it must consider the impacts of policy on demands to convert its currency and so must build into its policy some constraints to ensure that everybody believes convertibility on demand is possible. Generally, countries that tightly manage their exchange rates will formulate policy with a view to ensuring the accumulation of foreign currency or gold reserves, which can conflict with the pursuit of domestic policy goals such as full employment, growth and rising living standards.

Finally, a government with its own sovereign currency generally issues bonds denominated in that currency, servicing the debt with payments made in its own currency. However, it might choose to (or believe it must) issue debt in a foreign currency – in which case its policy space is constrained, as discussed above.

Taxes Drive Currency

A government that issues sovereign currency can choose the money of account, denominate taxes and other obligations such as fees and fines in that money of account and make and receive payments in its sovereign currency. By imposing taxes and other obligations on citizens payable in the sovereign's currency, government ensures there will be a demand for its currency. (For references to the idea that 'taxes drive money' in the history of thought, see Forstater 2005a, pp. 216–17.)

If citizens need currency to pay taxes, they will provide labour and other resources to government to obtain the currency. MMT argues that a tax payable in the currency is sufficient to drive the currency, since the citizens will want at least enough of it to pay taxes. (See Bell 2000; Tcherneva 2006.) Indeed, they will probably want more to save for the proverbial 'rainy day'. With a broad-based tax, the demand for currency will be generalised throughout the nation, leading to its use in third party transactions (within the nongovernment sectors).

Note that from inception government must provide the currency before taxes can be paid. Government can either spend or lend the currency into the economy. If government spends more than it taxes, it incurs a deficit with currency accumulating in the nongovernment sector. From inception, government cannot collect more currency in payment than it has issued. If over some period the government taxes more than it spends, the nongovernment sector dis-saves (runs down its currency balances accumulated during previous government deficits) or borrows currency from the government (goes into debt) to make the payments (with the government recording a budget surplus and accumulating credits against the nongovernment sector).

Sovereign Government Cannot Run Out of Its Own Currency and Cannot Be Revenue Constrained

The currency issuer cannot run out of currency that it spends or lends. Even if government does peg to gold or a foreign currency, it cannot run out of its own currency – but it can run out of whatever reserve it has promised for conversion. This is why countries that do not float their currencies can face diminished domestic policy space. Those that do float have more space, although they still might worry about the impacts of their spending (and lending) on exchange rates and on domestic inflation.

More generally, currency issuers can face resource constraints – as they hire labour and make purchases, they move resources to the government sector. At some point they begin to compete with private users of these resources, and can set off a bidding war that pushes up prices and wages. The danger of too much spending is inflation – not that government will run out of money to spend. In such a circumstance, government can either cut its spending or raise taxes (to push the nongovernment sector to reduce its spending) to prevent inflation.

Sovereign Government Does Not Face Solvency Risk in Its Own Currency and Cannot Be Forced into Involuntary Default

The currency issuer cannot be forced to default on its commitment to make a payment in its own currency. If government promises to pay wages, make a social security payment or pay interest in its own currency, it can always make that payment by issuing its currency. It might choose to default on its promise, but it cannot be forced to default against its will. However, a government that promises to convert its currency to foreign currency or gold, or that borrowed in foreign currency, might be forced to default if it cannot obtain sufficient reserves to meet the demand for conversion.

Sovereign Government Does Not Need to Borrow Its Own Currency in Order to Spend

If sovereign government can issue currency to finance its spending, why does it sell bonds? First, it is clear that government doesn't have to sell bonds so long as its currency is accepted in payment. It is hard to conceive of a plausible situation in which a modern developed nation would offer its currency in payment but find no domestic takers (while perhaps slightly more plausible, it is also difficult to believe that even foreign sellers would turn down currency, since they could take it to foreign exchange markets). The question, again, is whether currency-financed purchases might cause inflation and/or currency depreciation. In other words, the currency will be accepted, so the only question is about the price or exchange rate.

Second – and this gets a bit more technical – if one looks at the operational impact of bond sales, they essentially substitute one kind of government liability for another. Government accepts its own IOUs (technically, reserves, which are a part of high powered money – reserves plus cash – also called monetary base) in payment by buyers of another of its IOUs (government bonds). The main difference between reserves and bonds is that the latter pay higher interest rates. When government sells bonds, the nongovernment sector trades very short-term and low-interest-earning reserves for higher-earning and generally longer-term bonds. Note also that government is the source of both reserves and cash (reserves come from the central bank; typically, coins come from the treasury and paper notes come from the central bank) – and government must spend or lend those before the nongovernment sector can offer them in purchase of bonds. Hence the logic is that bonds are sold *after* government has already spent.

MMT sees sovereign bond sales as part of monetary policy operations rather than as a funding operation for government. By contrast, most economists see bond sales by the central bank (open market sales) as monetary policy, but sales of new issues by the treasury as part of fiscal policy. MMT argues that these are operationally identical, at least from the perspective of the nongovernment sector. Whether bonds are sold by the central bank or by the treasury, they reduce reserves held in the banks, which relieves downward pressure on interest rates. Bond sales – whether by the central bank or treasury – are an important lever that facilitates central bank interest rate targeting. (In the case where banks find themselves short of reserves – which pushes rates up – the government reverses policy, buying or retiring bonds and replacing them with reserves.)

Central Banks Are Never Really Independent

Central banks coordinate operations with the government's treasury. They act as the treasury's bank, making and receiving payments for government. They ensure that the treasury's cheques never bounce. Since they target interest rates, they cooperate with fiscal operations to ensure that when government spends or receives taxes, this does not affect bank reserve holdings in a way that would cause the interest rate to deviate from target. For this reason, they really cannot act independently from the treasury. Indeed, in times of unusual stress (such as major wars), the central bank is sometimes placed under the treasury's oversight.

For Every Surplus There Must Be a Deficit

A fundamental principle of macroeconomics accounting is that aggregate spending must equal aggregate income; all spending flows must go to someone as income. If one spends more than her income, another must spend less; if one sector of

the economy spends more than its income, another must spend less. MMT frequently divides the national economy into three sectors: domestic government (including national, state or province, and local), domestic private (including households, firms and not-for-profits) and foreign (foreign governments, firms, and households). While any sector can run a deficit (spend more than its income), that means that at least one of the other sectors must run a surplus (spend less than its income). Surplus sectors accumulate claims on deficit sectors; at the aggregate level, these claims (credits) accumulated by the surplus sectors equal the debts issued by the deficit sectors. If government runs a deficit, at least one of the other sectors (domestic private or foreign) must run a surplus; if government runs a surplus, at least one of the others must run a deficit.

Full Employment Is Affordable

The sovereign currency issuer can always afford to buy anything that a seller is willing to sell for that currency. This includes labour services. In other words, government can always afford to hire anyone who wants to work for wages paid in the government's currency. If the government does pursue a policy to maintain full employment, the potential dangers are rising wages and inflation, depreciation of the currency and removing labour resources from desirable uses in the private sector.

Non-affordability, i.e. running out of money, is not an issue for sovereign government and so is not a proper justification for policy choice. For example, social security cannot go bankrupt and solving the social security problem involves solving any real resource problems, not resolving (nonexistent) financial problems.

In the next section we provide a brief summary of the intellectual origins of MMT; in the final section we conclude with the most important policy implications.

The Foundations of Modern Money Theory

Modern Money Theory rests 'on the shoulders of giants', reviving the State Theory of Money (or Chartalism) and integrating it with a variety of heterodox approaches to macroeconomics, including the credit, circuitiste and endogenous approaches to money, the functional finance approach to budgeting, the financial instability hypothesis, and the sectoral balances approach. (See Mosler 2010; Wray 1998; Wray 2012/second edition 2015 for the MMT, state and endogenous money theories, as well as Graziani (1990) and Parguez and Seccarrecia (2000) for the circuit approach.) It draws heavily on historical, anthropological, sociological and legal interpretations of the nature of money, while also providing close analysis of modern monetary operations. (See Ingham 2005.) It is critical of orthodox approaches to fiscal and monetary theory and policy.

G. F. Knapp developed the State Theory of Money (published in German in 1905 and translated to English in 1924), building on Simmel's (1907) sociological approach to money (as well as the related German Historical approach). J. M. Keynes's *Treatise on Money* (1930) adopted Knapp's views on the role played by the state in choosing the *money of account* and in *enforcing* its use in payments:

> The State, therefore, comes in first of all as the authority of law which enforces the payment of the thing which corresponds to the name or description in the contracts. But it comes in doubly when, in addition, it claims the right to determine and declare what thing corresponds to the name, and to vary its declaration from time to time – when, that is to say, it claims the right to re-edit the dictionary. This right is claimed by all modern states and has been so claimed for some four thousand years at least. (Keynes 1930, p. 4)

He goes on to argue that 'Chartalism begins when the State designates the objective standard which shall correspond to the money-of-account' (Keynes 1930, p. 11). '[M]oney is the measure of value, but to regard it as having value itself is a relic of the view that the value of money is regulated by the value of the substance of which it is made, and is like confusing a theatre ticket with the performance' (Keynes 1983, p. 402). Money's 'substance' (whether stamped on coin or paper, or keystroked onto a computer's hard drive) is just a 'ticket', or record-keeping; but what is important

is that value is measured in terms of the money unit (including credits and debits denominated in that unit).

Keynes also seems to have been influenced by A. M. Innes, who independently integrated state and credit approaches to money in two articles published in a banking law journal (1913, 1914, reproduced in Wray 2004). Keynes actually reviewed the first of these articles in his *Economic Journal*, arguing that while some of the details might be subject to critique, the general argument appears correct (Keynes 1914). Of particular note was Innes's rejection of the typical approach to the origins of money – which supposes that money was created as a transactions cost-reducing innovation – and his speculation on money's true history, which can be traced through the innovation of measuring debt in a universal money of account. This seems to have led to what Keynes called his 'Babylonian Madness' – a period during which he explored the history of the earliest known monetary units, showing that they were always based on a specific number of grains of wheat or barley (Keynes 1983, pp. 233–6; see Ingham 2004, for a discussion of this period and for the creation of an abstract measuring unit). What this meant – for Knapp, Innes and Keynes – is that the money of account likely originated for official record-keeping purposes of debts and payments rather than evolving 'naturally' out of exchange based on barter.

Innes called his approach the 'Credit Theory of Money' and opposed it to the 'Metallic Theory', 'which has hitherto been held by nearly all historians and has formed the basis of the teaching of practically all economists on the subject of money'. More recently, Goodhart (1998) likewise distinguished between 'M' (metal, or Monetarist) form and 'C' (Cartalist or Chartalist) form, arguing that while the former can (still) count far more proponents, it is the latter that is supported by historical and anthropological evidence. The great numismatist Philip Grierson (1977) argued that the money of account probably developed out of the ancient tribal practice of *wergild* – imposing fines on transgressors, payable to victims, to prevent blood feuds – which emphasises the role played by authorities in choosing the nominal measuring unit and in enforcing obligations. Later, the system evolved to one in which obligations are to the authorities, denominated in the generalised money of account.

For application of this idea to the use of taxes to drive money in Africa, see Forstater (2005b, pp. 62–3):

> Direct taxation was used to force Africans to work as wage laborers, to compel them to grow cash crops, to stimulate labor migration and control labor supply, and to monetize the African economies. Part of this latter was to further incorporate African economies into the larger emerging global capitalist system as purchasers of European goods. If Africans were working as wage laborers or growing cash crops instead of producing their own subsistence, they would be forced to purchase their means of subsistence, and that increasingly meant purchasing European goods, providing European capital with additional markets. It thus also promoted, in various ways, marketization and commoditization. We have also seen that taxation was related to a variety of ideological aspects related to the reproduction of colonial relations of production. Direct taxation was thus an important 'secret of colonial capitalist primitive accumulation.' It appears to have been one of the most powerful policies in terms of its wide variety of functions, its universality in the African colonial context, and its success in achieving its intended effects. Of course, taxation was not the sole determinant of primitive accumulation. But it has certainly been under-recognized in the literature on primitive accumulation. The history of direct taxation in colonial capitalism also has some wider theoretical implications. It shows, for example, 'that "monetization" did not spring forth from barter; nor did it require "trust" – as most stories about the origins of money claim' (Wray 1998, p. 61). In the colonial capitalist context, money was clearly a 'creature of the state.'

J. Schumpeter (1934) distinguished between a 'Money Theory of Credit' and a 'Credit Theory of Money'. The first sees private 'credit money' as only a temporary substitute for 'real money'. Final settlement must take place in real money, which is the ultimate unit of account, store of value and means of payment. Exchanges might take place based on credit, but credit expansion is strictly constrained by the quantity of real money. Ultimately, only the quantity of real money matters so far as economic activity is concerned.

Most modern macroeconomic theory is based on the concept of a deposit multiplier that links the

quantity of privately created money (mostly bank deposits) to the quantity of high-powered money (HPM, which includes central bank reserves plus currency). This is the modern equivalent to Schumpeter's monetary theory of credit, and Friedman (or Brunner) is the best representative. In that view, the real money that is the basis of deposit expansion should be controlled, preferably by a rule that will make the modern fiat money operate more like the metallic money of the hypothesised past. (See Samuelson (1973) for the classic story of money's origins, from barter through commodity money and finally to fiat money.)

The credit theory of money, by contrast, emphasises that credit normally expands to allow economic activity to grow. This new credit creates claims on HPM even as it leads to new production. However, because there is a clearing system that cancels claims and debits without use of HPM, credit is not merely a temporary substitute for HPM. Schumpeter does not deny the role played by HPM as an ultimate means of settlement; he simply denies that it is required for most final settlements. The modern exponents of the credit theory of money include the Franco-Italian circuit approach (Graziani 1990) as well as the post-Keynesian endogenous money theory (Moore 1988; Wray 1990). Circuitistes envision a production process that begins with a bank loan to a firm, which hires labour to produce commodities. Banks create deposits credited to the accounts of firms as they make the loans, which are then transferred to workers as wages are paid. When households purchase output using their wages, the deposits are returned to firms, which can repay loans – closing the circuit as the deposits are debited.

Extensions to the theory have been made to allow for generation of profits and for payment of interest. For the MMT perspective on the monetary circuit, see Mosler et al. (1999, p. 177):

> This paper outlines an alternative way of viewing the monetary circuit that takes into consideration the central role of the State from the beginning of the analysis. Vertical and horizontal components of the monetary circuit were introduced and their relation analyzed. It was shown that this framework is applicable not only to currency, but to any commodity. This is because, while currency does not obtain its value by virtue of its status as a commodity, once endowed with value a tax driven currency can be analyzed like any other commodity.

The distinction between money and credit is not accepted by MMT, which is aligned with the credit theory. All monetary instruments are credit instruments, even HPM, because all of them represent promises of their issuer. Government promises to accept its currency in payments owed (taxes, fees and fines), and banks promise to accept their own monetary IOUs in payments owed to them. (They also promise to convert some of their liabilities to HPM on demand or after some waiting period – see below.) In fact, anybody can create a monetary instrument – i.e. create legal promises that are of a monetary nature – the problem is to get them accepted, as Minsky (1986) argued.

The endogenous money approach has focused on the implications of bank credit creation for supposed control by the central bank of the money supply. Since banks need reserves for clearing among one another and in some cases (such as in the USA) to meet legally required reserve ratios, the central bank normally accommodates the demand for reserves. The consequences of a central bank refusal to provide needed reserves are twofold: the central bank would lose control of the overnight interest rate and the smooth operation of the clearing system would be jeopardised. Hence, in practice, central banks always accommodate bank demand for reserves, albeit at the policy interest rate chosen by the central bank. In other words, central banks operate with interest rate targets, not reserve or money supply targets.

In recent years, this view has become accepted by most monetary policy-makers, even if the old 'deposit multiplier' is still presented in textbooks (Sheard 2013).

Post-Keynesians summarise the money creation process as 'loans create deposits' (banks create deposits in their loan-making activity) and 'deposits create reserves' (the central bank always accommodates the demand for expansion of the supply of reserves as needed when deposits

grow). This effectively reverses the logic of the textbook money multiplier exposition, even as it deals a fatal blow to the Monetarist policy rule that would have the central bank grow the money supply at a constant rate. Instead, the money supply grows as banks accommodate the demand for loans, with monetary policy-making consisting of setting the overnight rate target (which might, or might not, indirectly affect money growth).

Lerner (1943, 1947) developed the functional finance approach to sovereign government budgeting in the early post-war period to counter the belief of 'sound finance' that government ought to try to balance its budget. He posed two principles: the government's budget should balance only at full employment; if there is unemployment, government ought to spend more or reduce taxes; and (2) government ought to offer more bonds only if nongovernment sectors want to hold less money; if nongovernment sectors want to hold more money, government should reduce the supply of bonds. The first principle concerns the goal of fiscal policy – which is to set the budget so as to achieve full employment and without regard to whether that means a budget deficit (or balance, or surplus) will result. The second sets the monetary policy goal, which is to ensure that bank reserves are consistent with hitting the interest rate target. However, Lerner has integrated fiscal and monetary policy because the second principle would require that budget deficits are 'money financed' rather than 'bond financed' if the nongovernment sector's portfolio preferences are biased toward holding 'money'.

This is similar, although not identical, to a proposal by Friedman (1948), which would have the government finance all spending by issuing money that is returned when taxes are paid. A budget deficit would lead to net money creation, while a budget surplus would reduce the outstanding money supply. Friedman proposed that a balanced budget should be achieved only at full employment – with countercyclical deficits ('fiscal policy') as well as countercyclical movement of the money supply ('monetary policy') combining as powerful automatic stabilisers.

Beardsley Ruml (a New Dealer, Chairman of the Federal Reserve Bank of NY during the Second World War, and the 'father' of income tax withholding) argued that taxes had become 'obsolete' as a source of revenue for the federal government, and instead should be used to stabilise the purchasing power of the dollar (among other purposes, including income redistribution) (Ruml 1946a, b). According to Ruml, the budget deficits of the Second World War had shown that government spending is not revenue-constrained, but rather should be limited only when it threatens to spark inflation, and that a central bank can keep interest rates as low as desired no matter how large the government debt ratio becomes.

The similarities among these arguments, advanced by individuals with quite diverse perspectives, seems to indicate that such views were pervasive in the early post-war period – although by the 1970s the economics profession had returned to more conventional views on government finance (likening the government's 'budget constraint' to that of a household), and by the 1980s policy-makers had come to see budget deficits as problematic.

Minsky (1986) taught that the early post-war period was stabilised by the growth of 'big government' (the US government's share of GDP rose from 3% at the time of the Great Depression to an average of 20–25% in the post-war period) and the 'big bank' (a more active Fed intervening to stabilise interest rates and as lender of last resort). In his view, a large outstanding government debt actually promotes financial stability because it fills private portfolios with safe earning assets. However, he warned that the relative stability would (eventually) promote greater risk-taking as memories of the calamity of the 1930s faded; as he put it, 'stability is destabilizing'. Hence, behavioural changes in the private sector would gradually transform the financial structure from 'hedge' (the safest, where income flows are expected to be sufficient to make all payments as they come due), to 'speculative' (where only interest could be paid – principal could not be retired), and finally to 'Ponzi' (interest would have to be capitalised – the debtor borrows to pay interest). Financial crises would reappear and become more frequent and more severe.

Minsky argued that expansions that are led by government spending, and in which consumption is financed out of income flows generated in a high-employment society, are more sustainable than are expansions led by private sector investment or debt-fuelled borrowing. Godley (1996) developed a simple but highly instructive sectoral balances approach along a similar vein. At the aggregate level, if one sector runs a deficit (spending more than its income), then by identity at least one other sector must be running a surplus (spending less than income, and accumulating claims on the deficit sector). In a closed economy, the private sector can run a surplus ('saving') only if the government sector runs a deficit; the debt of the government sector equals the net accumulation of financial assets ('net financial saving') of the private sector. Allowing for a current account deficit means that the government's budget deficit needs to be larger – the sum of the current account deficit and the private sector surplus will equal the budget deficit. This adds weight to the arguments of Minsky, Lerner and Ruml that pursuit of a balanced budget can be counterproductive – especially for a nation like the USA that runs trade deficits. A balanced budget policy would mean, by identity, that the USA's private sector would run chronic deficits and dig itself deeper into debt. (This is precisely what happened in the decade leading up to the global financial crisis – see Godley and Wray (1999) as well as Tymoigne (2014a).)

Most of these traditions were largely lost over the final four decades of the twentieth century. The Chartalist approach to money never gained much of a foothold, with economics textbooks continuing to propound the barter approach to money. Indeed, money became increasingly unimportant in sophisticated economic theory – at most, serving to lubricate market exchange, and having little to no 'real' impact in most macroeconomic models. In the late 1960s, microeconomic theory of the consumer's budget constraint was adapted as a government budget constraint – with government choosing to finance its spending through tax revenue, borrowing by issuing bonds or printing money. Borrowing could drive up interest rates (crowding out investment) and expose government to default risk; printing money raised the spectre of inflation (or even hyperinflation). Minsky's warnings of the potential for financial crisis were largely ignored; indeed, by the 1990s and 2000s, the 'era of the Great Moderation' had supposedly arrived, a period of diminished risk and greater stability. Godley's sectoral balance approach never gained many adherents – while economists and policy-makers clamoured for government budget surplus *and* simultaneously for more private saving (an impossible combination except for current account surplus nations). The only stream of research that did make headway was the endogenous money approach, as policy-makers abandoned money targets in favour of interest rate targeting (albeit in the guise of a Taylor rule to control inflation).

However, MMT used these foundations to build a robust approach to macroeconomics. Beginning in the mid-1990s, MMT developed a large following over the next two decades especially after the growth of the 'blogosphere', which helped to spread the ideas outside academia. MMT's standing was also increased by the global financial crisis after 2007 as well as the entrenched euro crisis that began a couple of years later because its proponents had long warned of the likelihood of each (for an early warning, see Godley 1992; for analysis of the crisis see Wray 2009, 2012).

We will explore the main policy implications in the following section.

Policy Implications of MMT

MMT follows in the tradition of those early post-war economists who recognised that sovereign government cannot run out of its own currency. As such, government faces an inflation constraint, not a solvency constraint. MMT also adopts the Knapp–Innes–Keynes–Lerner view that the state chooses the money of account, imposes taxes and other obligations in that unit, and issues the currency that it accepts in payment of those obligations. For that reason, government must spend first before it can collect taxes. Government's net (deficit) spending allows the nongovernment sector to run a surplus (net financial saving), accumulating claims on government. These claims can

take the form of currency, bank reserves held at the central bank or bonds issued by the treasury.

If government floats its currency, it maximises its domestic fiscal and policy space. It can formulate its spending and tax policy to pursue domestic policy goals such as full employment, price stability and rising living standards. It can also set its interest rate target consistent with those goals and with achieving a desired distribution of income between creditors and debtors. Government might instead choose to manage its exchange rate along a continuum between loosely held ceilings and floors on one end to a tightly held peg with a promise to convert on demand at a fixed exchange rate at the other end. A managed exchange rate regime reduces domestic policy space; a peg opens the possibility of default on the promise to convert and exposes the government to speculative attacks and currency crises.

MMT has made major contributions to our understanding of the coordination of fiscal and monetary policy. In modern nations, the monetary and fiscal policy functions are divided between the central bank and the treasury. While it is often claimed that central bank independence is desirable (the claim is that this enhances the central bank's ability to fight inflation, as it can supposedly refuse to allow the treasury to 'money finance' spending), in fact the central bank and treasury must closely coordinate their operations. The modern central bank makes and receives payments for the treasury, clearing payments between the treasury and private banks (while also clearing payments among private banks and acting as lender of last resort). Since central banks operate with an overnight interest rate target, and since these daily payment flows are huge, the central bank always accommodates payment system needs for reserves. For a discussion of the necessity of central bank intervention to keep rates on target, see Forstater and Mosler (2005, p. 539) who argue:

> In a state money system with flexible exchange rates running a budget deficit – in other words, under the 'normal' conditions or operations of the specified institutional context – without government intervention either to pay interest on reserves or to offer securities to drain excess reserves to actively support a nonzero, positive interest rate, the natural or normal rate of interest of such a system is zero.

In modern systems, the central bank either pays interest on reserves or offers interest-paying treasury bonds as an alternative to reserves. Since the treasury is by far the largest economic entity in any modern economy, its fiscal operations have huge impacts on daily payments flows that must be offset by central bank operations. So while the central bank may have substantial discretion in choosing its interest rate target, its monetary operations cannot be formulated independently from treasury operations (Fullwiler 2011; Tymoigne 2014b).

MMT recognises the symmetry between the post-Keynesian view that bank loans 'create' deposits, which then leads to the creation of reserves as the central bank accommodates demand, and the Chartalist view that government must spend before it can collect taxes. Bank deposits are the liabilities of banks, and must be created when the bank makes a loan; central bank reserves are the liabilities of the central bank and must be created when the central bank either lends reserves to banks, or purchases assets (such as government bonds) from them; and currency is the liability of the government that must be created by government as it spends. Only once deposits are created can debtors to banks use them to make payments; only after reserves are created can banks use them to repay loans to the central bank, or use them to buy government bonds; and only after currency has been created can taxpayers use it to pay taxes and other obligations to the state. Today these operations occur on balance sheets as electronic entries and debits. Banks cannot run out of deposits; central banks cannot run out of reserves; and sovereign governments cannot run out of currency. This symmetry is hidden in most analyses of fiscal policy behind the veil of central bank independence and government budget constraints. If economists understood the coordination of monetary and fiscal policy operations, they would understand Ruml's claim that 'taxes for revenue are obsolete'.

Another area on which MMT has focused is policy to achieve and maintain full employment with wage and price stability. Following Minsky (1965), MMT adopted the 'employer of last resort'

or 'job guarantee' proposal in which the national government provides wages to fund a programme that would offer a job to anyone who wants to work (Tcherneva 2014; Mitchell and Wray 2005; Mitchell and Muysken 2008; Harvey 1989; Forstater 1999; Wray and Forstater 2004; Mosler 1997–98.) The universal job guarantee was part of Martin Luther King's proposal to reduce inequality and 'depression-like' unemployment suffered by African Americans even during the business cycle upswing of the 1960s. For example, see Forstater (2002, p. 45), who argues:

> The Rev. Dr. Martin Luther King, Jr. wrote extensively on economic matters, especially unemployment policy. King supported a federal job guarantee for anyone ready and willing to work. He believed it would provide employment and income security, as well as increased public and community services. . . His policy proposals are just as relevant today as they were when they were first put forward some forty years ago.

While there are various versions of the proposal, most of MMT's followers advocate a universal programme that would offer a uniform basic wage plus benefit package, taking workers 'as they are' and 'where they are' – creating jobs in every community and tailoring them to the skills and educational level of workers. The programme could be decentralised, with projects formulated and managed locally (by local governments, school and park districts and not-for-profit community service organisations), but with funding from the central sovereign government (the only entity that can afford an open-ended offer to hire anyone who wants to work). The projects would provide skills and training upgrading on the job and would generate output useful to the community.

Unlike for-profit business, because the programme would not have to operate according to profit-maximisation criteria, it could pursue other goals, such as the creation of 'green jobs' and promoting social, environmental, economic and financial stability. It would also provide a powerful automatic stabiliser – with government spending on wages in the programme rising when the private sector slows, and falling when the private sector heats up. Private employers would recruit from the programme's pool of labour, which would act like a buffer stock to help stabilise wages (and hence prices). MMT's proponents argue that this 'reserve army of the employed' will function much more effectively than a Marxian 'reserve army of the unemployed', as it would allow workers to preserve and even upgrade their skills rather than becoming unemployed whenever the private sector downsizes. In a slump, the programme would prevent wages from falling below the programme's wage, helping to stabilise consumption. In this way, the programme helps to maintain the advantages of private market flexibility even as it maintains full employment and reduces fluctuation of aggregate demand. As Forstater (1998, pp. 562–3) put it:

> Full employment and even high employment and capacity utilization rates are associated with structural rigidities related to a number of undesirable consequences. For this reason, central banks, national governments, and international organizations have resisted policies that would promote full employment. What has been almost entirely overlooked, however, is the ways in which the selective use of discretionary public employment might promote higher levels of employment without the loss of system flexibility. A primary reason for overlooking the advantages of public employment has been due to the tendency to evaluate public sector activity by the same criteria that private sector activity is evaluated. But public sector activity serves a different purpose than private sector activity and so should be evaluated according to different criteria. The public sector is not constrained by the same competitive pressures as the private sector, and therefore it has a greater degree of latitude in choosing what activities to engage in, what methods of production to utilize, and where to locate its activities. These characteristics of public sector activity may be utilized to promote higher levels of employment without resulting in rigidities of the production system normally associated with high or full employment. In addition, these same features may also enable these higher levels of employment without undesirable environmental impacts or geographic dislocation of workers.

MMT was an early critic of the setup of the European Monetary Union, arguing that the attempt to divorce monetary policy formation (in the hands of the ECB) from fiscal policy (which remained within the purview of the individual member nations) was a mistake. More specifically – as Goodhart (1998) put it – the EMU was the first major deviation from the 'one

nation, one currency' rule that we find going back through history and around the globe today. The creation of the EMU was a conscious attempt to eliminate what we have called here sovereign currency, with each member nation adopting what was essentially a foreign currency. As currency users (not issuers), spending by member governments would be limited to their tax revenue plus ability to borrow euros. They became somewhat analogous to US states or Canadian provinces – ultimately relying on the willingness of the ECB to stand behind their governments' debts.

MMT followers warned that the first serious financial crisis and deep recession would cause budget deficits to rise, triggering credit downgrades and higher interest rates on debt. Facing default, member nations had to turn to the 'Troika' which imposed austerity as a condition of lending. As Godley (1992) had feared, loss of their own fiscal sovereignty would reduce members to the status of colonies. By 2015, these fears had been validated – at least in Greece. Many came to see the solution to the euro crisis as requiring reconnection of fiscal policy and currency sovereignty, either for the union as a whole or through dissolution of the EMU and restoration of national moneys (Mitchell 2015).

Conclusion

MMT has synthesised a number of themes that can be traced back through the history of economic thought, but which had largely been lost over the past half century as economics turned toward increasingly sophisticated mathematicised models that downplay the role of money and financial institutions. In place of detailed analysis of the coordination of monetary and fiscal operations, most economics supposed 'money drops' and 'government budget constraints'. The importance of fiscal sovereignty was ignored – or dismissed – as many pushed for 'optimal currency areas', currency boards and dollarisation or euroisation. Rather than considering the historical evolution of money and financial institutions, simplistic stories of the transition from barter to commodity money focused attention on money's role as a medium of exchange. While that is, of course, an important function of money, it has little to do with most of the financial innovations that led up to the global financial crisis after 2007. MMT has tried to recover the messier but more revealing traditions that have survived on the fringes of the discipline. Much remains to be done to make monetary economics relevant to the real world.

See Also

- ▶ Functional Finance
- ▶ Historical School, German
- ▶ Knapp, Georg Friedrich (1842–1926)
- ▶ Lerner, Abba Ptachya (1905–1982)
- ▶ Minsky Crisis
- ▶ Quantity Theory of Money
- ▶ Schumpeter, Joseph Alois (1883–1950)

Bibliography

Bell, S. 2000. Do taxes and bonds finance government spending? *Journal of Economic Issues* 34: 603–620.

Forstater, M. 1998. Flexible full employment: Structural implications of discretionary public sector employment. *Journal of Economic Issues* 32(2): 557–563.

Forstater, M. 1999. Full employment and economic flexibility. *Economic and Labour Relations Review* 11: 69–88.

Forstater, M. 2002. 'Jobs for all': Another dream of the Rev. Dr. Martin Luther King, Jr. *Forum for Social Economics* 31(2): 45–53.

Forstater, M. 2005a. Tax-driven money: Additional evidence from the history of thought, economic history and economic policy. In *Complexity, endogenous money, and exogenous interest rates*, ed. M. Setterfield, 202–220. Cheltenham: Edward Elgar.

Forstater, M. 2005b. Taxation and primitive accumulation: The case of colonial Africa. *Research in Political Economy* 22: 51–64.

Forstater, M., and W. Mosler. 2005. The natural rate of interest is zero. *Journal of Economic Issues* 39(2): 535–542.

Friedman, M. 1948. A monetary and fiscal framework for economic stability. *The American Economic Review* 38(3): 245–264.

Fullwiler, S. 2011. Treasury debt operations: An analysis integrating social fabric matrix and social accounting matrix methodologies. SSRN. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1825303

Godley, W. 1992. Maastricht and all that. London Review of Books. Available at: http://www.lrb.co.uk/v14/n19/wynne-godley/maastricht-and-all-that

Godley, W. 1996. Money, finance and national income determination: An integrated approach. Levy Economics Institute. Working Paper 167, June. Available at: http://www.levy.org/

Godley, W., and L.R. Wray. 1999. Can Goldilocks survive? Policy Note 1999/4, April, Levy Economics Institute. Availability at: http://www.levyinstitute.org/pubs/pn99_4.pdf

Goodhart, C.A.E. 1998. Two concepts of money: Implications for the analysis of optimal currency areas. *European Journal of Political Economy* 14: 407–432.

Graziani, A. 1990. The theory of the monetary circuit. *Economies et Societes*, series no. 7, June.

Grierson, P. 1977. *The origins of money*. London: Athlone Press.

Harvey, P. 1989. *Securing the right to employment: Social welfare policy and the unemployed in the United States*. Princeton: Princeton University Press.

Ingham, G. 2004. *The nature of money*. Cambridge: Polity Press.

Ingham, G., eds. 2005. *Concepts of money: Interdisciplinary perspectives from economics, sociology, and political science*. Cheltenham: Edward Elgar.

Innes, A.M. 1913. What is money? *Banking Law Journal*, May, pp. 377–408

Innes, A.M. 1914. The credit theory of money. *Banking Law Journal*, January, pp. 151–168; Reprinted in *Credit and state theories of money*, ed. L.R. Wray, pp. 14–49, 2004. Cheltenham/Northampton: Edward Elgar.

Keynes, J.M. 1914. What is money? *Economic Journal* 24(95): 419–421.

Keynes, J.M. 1930. *A treatise on money*, Vols. I and II (1976). New York: Harcourt, Brace & Co.

Keynes, J.M. 1983. *The collected writings of John Maynard Keynes*, ed. D. Moggridge, Vol. XXVII-I. London/Basingstoke: Macmillan.

Knapp, G.F. 1924. 1973. *The state theory of money*. Clifton: Augustus M. Kelley.

Lerner, A.P. 1943. Functional finance and the federal debt. *Social Research* 10: 38–51.

Lerner, A.P. 1947. Money as a creature of the state. *American Economic Review* 37: 312–317.

Minsky, H.P. 1965. The role of employment policy. In *Poverty in America*, ed. M.S. Gordon. San Francisco: Chandler Publishing Company.

Minsky, H.P. 1986. *Stabilizing an unstable economy*. New Haven/London: Yale University Press.

Mitchell, W. 2015. *Eurozone dystopia: Groupthink and Denial on a grand scale*. Cheltenham: Edward Elgar.

Mitchell, W., and J. Muysken. 2008. *Full employment abandoned: Shifting sands and policy failures*. Cheltenham/Northampton: Edward Elgar.

Mitchell, W.F., and L.R. Wray. 2005. In defense of employer of last resort: A response to Malcolm Sawyer. *Journal of Economic Issues* 39(1): 235–245.

Moore, B.J. 1988. *Horizontalists and verticalists: The macroeconomics of credit money*. Cambridge: Cambridge University Press.

Mosler, W. 1995. *Soft currency economics*. 3rd ed. West Palm Beach: III Finance.

Mosler, W. 1997. Full employment and price stability. *Journal of Post Keynesian Economics* 20(2): 167–182.

Mosler, W. 2010. *The seven deadly innocent frauds of economic policy*. US Virgin Islands: Valence Co., Inc.

Mosler, W., and M. Forstater. 1999. A general framework for the analysis of currencies and commodities. In *Full employment and price stability in the global economy*, ed. P. Davidson and J. Kregel, 166–177. Cheltenham: Edward Elgar.

Parguez, A., and M. Seccarrecia. 2000. The credit theory of money: The monetary circuit approach. In *What is money?* ed. J. Smithin, 101–123. London/New York: Routledge.

Ruml, B. 1946a. Taxes for revenue are obsolete. *American Affairs* 3(1), 35–39. Available at: http://www.constitution.org/tax/us-ic/cmt/ruml_obsolete.pdf

Ruml, B. 1946b. Tax policies for prosperity. *The Journal of Finance* 1(1): 81–90.

Samuelson, P. 1973. *Economics*. 9th ed, 274–276. New York: McGraw-Hill.

Schumpeter, J.A. 1934. *The theory of economic development: An inquiry into profits, capital, credit, interest and the business cycle*. Cambridge: Harvard University Press.

Sheard, P. 2013. *Repeat after me: Banks cannot and do not 'lend out' reserves*. New York: Standard & Poor's, Credit Market Services, Global Economics and Research. Available at: https://www.kreditopferhilfe.net/docs/S_and_PRepeat_After_Me_8_14_13.pdf

Simmel, G. 1907 (transl. 1978). *The philosophy of money*. Routledge Reprint 2011. London: Routledge.

Tcherneva, P. 2006. Chartalism and the tax-driven approach to money. In *Handbook of alternative monetary economics*, ed. P. Arestis and M. Sawyer, 69–86. Northampton: Edward Elgar.

Tcherneva, P. 2014. Reorienting fiscal policy: a bottom up approach. *Journal of Post Keynesian Economics* 37(1): 43–66.

Tymoigne, E. 2014a. A financial analysis of monetary systems. In *Contributions to economic theory, policy, development and finance*, ed. D.P. Papadimitriou, 88–113. New York: Palgrave MacMillan.

Tymoigne, E. 2014b. Modern money theory and the interrelation between the treasury and the central bank: The case of the United States. *Journal of Economic Issues* 48(3): 641–662.

Wray, L.R. 1990. *Money and credit in capitalist economies: The endogenous money approach*. Aldershot/Brookfield: Edward Elgar.

Wray, L.R. 1998. *Understanding modern money: The key to full employment and price stability*. Northampton: Edward Elgar.

Wray, L.R., eds. 2004. *Credit and state theories of money: The contributions of A. Mitchell Innes*. Cheltenham: Edward Elgar.
Wray, L.R. 2009. The rise and fall of money manager capitalism: A Minskian approach. *Cambridge Journal of Economics* 33(4): 807–828.
Wray, L.R. 2012. *Modern money theory: A primer on macroeconomics for sovereign monetary systems.* Basingstoke: Palgrave Macmillan (2nd edition forthcoming 2015).
Wray, L.R., and M. Forstater. 2004. Full employment and economic justice. In *The institutionalist tradition in labor economics*, ed. D. Champlin and J. Knoedler, 253–272. Armonk: M. E. Sharpe.

Modigliani, Franco (1918–2003)

Richard Sutch

Abstract

This article focuses on the scholarly contributions of Franco Modigliani, 1985 Nobel laureate in economics. Particular attention is given to his formulation of the determinants of equilibrium in Keynesian macroeconomics, the life-cycle hypothesis of saving, his contributions to the theory of expectations, and the Modigliani–Miller theorems of corporate finance. The objective is to demonstrate Modigliani's importance in the history of economics.

Keywords

Aging and retirement; American Economic Association; American Finance Association; Ando, A.; Arbitrage; Barro, R.; Bequest motive; Bonds; Budget deficits; Buffer stocks; Business cycle; Classical economists; Consumption function; Corporate finance; Cost of capital; Dividend policy; Duesenberry–Modigliani hypothesis; Econometric Society; Expectations; Federal Reserve System; Finance theory; Fiscal policy; FMP model; Friedman, M.; Great Depression; Greenspan, A.; Hicks, J.; Life cycle hypothesis; Life-cycle theory of saving; Infinite horizons; Inflation; Information cost; Inventories; Investment function; IS–LM model; Keynes, J. M.; Keynesian revolution; Koopmans, T. C.; Lange, O. R.; Lerner, A. P.; Life-cycle hypothesis; Liquidity trap; Long-term contracts; Marginal utility analysis; Marschak, J.; Microfoundations; Miller, M.; MIT model; Modigliani; Franco; Modigliani–Miller theorem; Monetarism; Monetary policy; Money demand; Money income vs real income; Money supply; Muth, J. F.; National debt; Neoclassical synthesis; New classical macroeconomics; Permanent income hypothesis; Precautionary savings; Preferred maturity habitat; Price control; quantity theory of money; Rational expectations; Real money; Relative income hypothesis; Ricardian equivalence theorem; Rules vs discretion in policy; Saving and growth; Saving–income ratio; Shiller, R.; Simultaneous equations; Social Security (USA); Stabilization policy; Sticky wages; Stock price volatility; Sutch, R.; Term structure of interest rates; Uncertainty; Underemployment equilibrium

JEL Classifications

B31

Franco Modigliani was awarded the Sveriges Riksbank (Bank of Sweden) Prize in Economic Sciences in Memory of Alfred Nobel in 1985 for 'pioneering studies of saving and of financial markets'. A life-long Keynesian, his contributions to macroeconomics and finance transformed both fields. The life-cycle approach to consumption and saving pioneered a microfoundations approach to macroeconomic theory and remains the standard model of consumption in macroeconomics. The Modigliani–Miller theorems on the cost of capital had a profound influence on subsequent research in finance. He was also a pioneer in modelling expectations in macroeconomic models. Modigliani was an influential and critical voice on macroeconomic policy in the United States, in his native country of Italy, and in the European community.

Biography and Intellectual Development

Modigliani was born in Rome, Italy on 10 June 1918. His father, who died when Modigliani was only 14, was a pediatrician. He entered the University of Rome to study law at 17. In his second year he won a national competition in economics with an essay on the price controls imposed in Italy during the annexation of Abyssinia (now Ethiopia). He records in his autobiography (2001) that, following the receipt of this award, he began a self-study of economics reading the classics, an approach he deemed more satisfactory then taking courses during the fascist regime.

At about the same time he became a committed anti-fascist. After the Italian government promulgated anti-Semitic laws in 1938, he and his fiancée, Serena Calabi, fled to Paris, where they were married in 1939. He and Serena applied for an immigration visa to the United States and arrived in New York in August 1939, a few days before the beginning of the Second World War. Modigliani was immediately taken on as a postgraduate scholar by the New School for Social Research, which had been newly created as a haven for social scientists fleeing Europe. He was mentored there in economic theory and econometrics by Jacob Marschak. Modigliani always took care to acknowledge the powerful influence that Marschak had on his development as an economist. During 1941–3 Modigliani taught as an instructor at the New Jersey College for Women (now Douglass College and at the time part of Rutgers University) and at Bard College of Columbia University (now independent). During these years he continued to work on his doctoral dissertation in social sciences for the New School, and received a Ph.D. in 1944. This work was reported in the same year in his first published article, 'Liquidity Preference and the Theory of Interest and Money'. He then returned to the New School as a lecturer.

Modigliani taught briefly at the University of Illinois (1949–52), where he was promoted from associate professor to full professor in 1950 at the age of 32. There he found a friend and collaborator, Richard Brumberg, a graduate student. Together they developed the life-cycle theory of saving, which became Modigliani's most important contribution and one of the two cited by the Nobel judges. He next taught at the Carnegie Institute of Technology (now Carnegie–Mellon) as a Professor of Economics and Industrial Administration (1952–60). Most of the other projects now associated with his name were begun there including his collaboration with Merton Miller on the founding theorems of corporate finance. These theorems were the second contribution cited in 1985 by the Nobel committee.

Modigliani visited Harvard University (1957–8) and the Massachusetts Institute of Technology (1960–1). He was appointed to the faculty at Northwestern University (1960–2) and taught there one year before returning to MIT in 1962 as a Professor of Economics and Finance. He remained at MIT for the balance of his career. By the mid-1960s MIT was regarded as the premier graduate school in the world for the study of economics.

Modigliani's scientific output is impressive for its breadth of coverage, the depth to which each topic was pursued, and the sheer volume of brilliant, highly original papers. In six volumes of collected papers (1980–2005), Modigliani assembled 87 published papers from a corpus of nearly 200 (and a famous previously unpublished paper with Richard Brumberg 1980). Modigliani also wrote or coauthored ten books and edited several more. This huge output is all the more remarkable when one considers that throughout his academic career Modigliani always subjected his economic theory to rigorous empirical verification, often employing sophisticated statistical technique with ingeniously (and laboriously) derived data.

In 1970 Modigliani was named an Institute Professor, an honorific title that MIT reserves for scholars of great distinction. He was elected President of the American Economic Association (1975–6). He also served as President of the Econometric Society and the American Finance Association. He became Professor Emeritus in 1988.

Franco Modigliani died on 25 September 2003 at the age of 85 in Cambridge, Massachusetts. MIT Institute Professor Paul Samuelson, a colleague and friend, said, 'Franco Modigliani

could have been a multiple Nobel winner. When he died he was the greatest living macroeconomist. He revised Keynesian economics from its Model-T, Neanderthal, Great Depression model to its modern-day form' (MIT 2003).

The Keynesian Revolution and the Debate Over Stabilization Policy

When he arrived in New York in 1939, Modigliani began several years of study of macroeconomics (and mathematics and statistics as well) under the tutelage of Jacob Marschak, Abba Lerner, Oskar Lange and Tjalling Koopmans. The hot topic, of course, was *The General Theory of Employment, Interest, and Money* by John Maynard Keynes. Published in 1936, Keynes's analysis was truly revolutionary. Keynes pioneered modern macroeconomics by proposing a novel and compelling explanation of the gyrations of the economic system. Those fluctuations had had a devastating impact on the US economy during the Great Depression of the 1930s, a catastrophe whose lingering effects were still evident in 1939. The Keynesian model also suggested a set of active policy prescriptions that might be used, first, to lift an economy out of depression and, second, to prevent recessions and depressions from occurring in the first place. Furthermore, Keynes suggested that if the curative policies were not applied, the economy might languish with mass unemployment for a long time.

Neither the theoretical formulation nor the policy prescriptions of *The General Theory* were easy to accept in the early 1940s. Keynes's argument was complicated and subtle, the book's prose was at points cumbersome and inelegant, and the concepts that Keynes introduced were unfamiliar to economists and sometimes counterintuitive. The policy implications seemed almost impossibly unorthodox. Government spending should not be based on the need for public services. Taxes should not be based on the need for revenue to pay for the government services. Instead government spending and taxation should be directed to restoring and then maintaining full employment which might lead to levels of spending far in excess of the perceived need for public services and to a level of taxation that might produce substantial deficits.

Working in 1942 and 1943, Modigliani sought to reduce the confusion generated by the debate over what Keynes was saying and to articulate the common sense of the Keynesian policy message. In the process he made an important clarification of the Keynesian argument. The result was his now famous *Econometrica* paper of 1944, 'Liquidity Preference and the Theory of Interest and Money'. The paper did three things. First, Modigliani reduced the 384 pages of Keynes's complex argument to a mathematical system of nine simultaneous equations. The virtue of a mathematical representation is that it served to insure that the variables considered important by Keynes were consistently and precisely defined and that the relationships among them were made rigorously explicit. Modigliani was not the first to attempt a mathematical reduction to clarify the logical structure of *The General Theory.* One of his mentors, Oskar Lange, the noted Polish economist, had preceded him. But Modigliani's version became the standard, taught to graduate students for decades (who generally left *The General Theory* unread), until it was replaced by a revised (and more complex) presentation produced by Modigliani in 1963, 'The Monetary Mechanism and its Interaction with Real Phenomena.' Second, Modigliani clarified the role played in the model by Keynes's assumption that money wages were inflexible. We return to this point below. Third, Modigliani argued that fiscal policy was not the only weapon available for fighting recessions. Monetary policy could be effective in many, if not all cases. In this third effort, Modigliani was taking issue with another of his mentors, Abba Lerner, who was suggesting at that time that fiscal policy, and only fiscal policy, would work.

The mathematical formulation of the determinants of macroeconomic equilibrium did much to make Keynes acceptable to economists, though it must be said that the mathematics required was most easily mastered by young economists still in graduate school or only recently accepted into the professorship. Many of the 'old guard' seemed

unable or unwilling to shed their pre-Keynesian conceptions. By salvaging and later defending a role for monetary policy, Modigliani had a major influence on the conduct of anti-recession policy, particularly in the 1950s and 1960s. But it was the clarification of the role of 'sticky wages' that helped to transform Keynesian economics into its modern form.

Modigliani established that both the classics and Keynes shared a conception of the macroeconomic demand for money derived from basic microeconomic principles. Any such model would necessarily connect money to real variables such as output and employment only when money entered the formulation as a ratio to the price level. This ratio is known as *real* money and was defined by Keynes and Modigliani in terms of 'wage units'. The 'classical' quantity theory of money, for example, made real money proportional to real output. If changes in the nominal money supply are to influence real output, there must be some reason why those changes are not immediately followed by an equi-proportionate change in wages. In the pre-Keynesian, 'classical' model wages would adjust rapidly, unemployment would thus be briefly transitory, and monetary policy would be both ineffective and unnecessary to increase output and employment.

Modigliani's equations revealed that idle resources and price flexibility could simultaneously exist only in the extreme case when the demand for money became infinite. Modigliani considered this an unlikely situation which he called the 'Keynesian case'. It later became better known as the 'liquidity trap'. In the *General Theory* Keynes had been critical of the flexible wage assumption and introduced what Modigliani considered the more realistic assumption that, in the short run at least, money wages would not adjust in the downward direction. With this specification added to the system of equations, underemployment equilibrium was possible even when liquidity trap conditions were not present. Modigliani argued on this basis that the hypothesis of wage rigidity was a necessary part of the Keynesian system if monetary policy was to play a role in influencing real variables.

As a corollary of the argument, Modigliani pointed out that the economy could not be 'dichotomized' into real and monetary sectors that operated independently of each other. In his demonstration Modigliani was following John Hicks who made the same point with the IS–LM apparatus made famous by introductory textbooks. In Hicks's (1937) diagram the interest rate and the level of output are jointly determined by the intersection of an LM curve reflecting an equilibrium of the demands and supplies that characterize the monetary sector (L for 'liquidity preference' and M for the money supply) with the IS curve reflecting the equilibrium of real forces (I for the demand for investment and S for the supply of saving). It might be noted, however, that Hicks expressed the IS–LM relationship in terms of the rate of interest and *money* income. It was Modigliani who gave it the appropriate interpretation in terms of the interest rate and *real* income.

Modigliani considered the 1944 paper one of his most significant contributions. It set the stage for the 'neoclassical synthesis' of the Keynesian and the classical traditions. This synthesis came to dominate the economics profession for the next three or four decades. That approach accepted that labour and capital would be underutilized over the course of the business cycle, that unemployment was not a transitory problem but a variable that helped clear the money market, and that activist monetary and fiscal policies can be welfare improving. Indeed, avoiding unemployment would take close management of the money supply and interest rates. In the United States these views became most influential during the 1960s when the administration of John F. Kennedy put them into practice in a serious way. But these academic and political developments pulled Modigliani into an extended debate with the 'monetarists'. Led by Milton Friedman, the monetarists held that the quantity of money is the key factor in determining economic change and that the fiscal variables advocated by Modigliani and other Keynesians are not important. Modigliani was particularly disturbed by Friedman's proposal that neither discretionary fiscal nor monetary policy should be employed; rather, the money supply should be strictly regulated to grow at a constant

rate (say three per cent per year). Modigliani ridiculed this prescription as a 'blind rule' and consistently argued that wise discretionary control of the money supply was essential.

In a pair of empirical papers, one with Albert Ando, his student at Carnegie, Modigliani, went on the attack (1964, 1965). In his presidential address to the American Economic Association, Modigliani rejected the idea that Keynesians did not think that money mattered, and he cited his 1944 and 1963 papers as proof (1977). After winning the rhetorical and empirical debate with Friedman, he sought to 'make peace' with the monetarists by declaring 'We are all monetarists'. And yet he went on to defend the case for policy discretion in a fashion he later described as 'a full, passionate, and polemical'. In an interview conducted in 1999, he declared victory. 'There is not a country in the world today that uses a mechanical rule' (2000, p. 236). He might have added that the highly praised success of Alan Greenspan as the Chairman of the U.S. Federal Reserve Board was based on the careful discretionary management of money and interest rates. In a series of lectures, later published as *The Debate over Stabilization Policy* (1986c), Modigliani traced the history of these disputes.

The monetarist debates of the 1960s, and the empirical success of the work testing Keynesian propositions, led to another important direction for Modigliani's research. He was asked to construct an econometric model of the US economy by the Federal Reserve. The model would be an empirically estimated system of simultaneous equations that would be used by the Federal Reserve to make and guide policy and forecast future developments. He asked Albert Ando to join him on the project and they created what was first known as the 'MIT model' and, after Ando moved to the University of Pennsylvania, as the 'Federal Reserve-MIT-University of Pennsylvania Model' (FMP) (1975a). The result embodied many of Modigliani's ideas about the structure of the economy, the consumption function, the structure of interest rates, and the workings of other financial markets. Emblematic of Modigliani's willingness to learn from the data were the many modifications he made to his early formulations in the process of constructing the FMP model. In particular he explicitly extended the theories to include the causes and consequences of inflation which had only begun to become a noticeable problem for the American economy in the 1970s (for his major contributions on inflation see Part III of *Collected Papers*, vol. 5). The model proved sufficiently valuable that the Federal Reserve continued to use it into the 1980s.

The Life-Cycle Model of Saving and Consumption

In his 1944 paper on the Keynesian model, Modigliani presented an equation for the national flow of saving that described saving as a positive function of aggregate income in a manner consistent with the 'consumption function' famously introduced by John Maynard Keynes in the *General Theory*. Keynes had postulated a 'fundamental psychological law' whereby an individual's consumption would increase as his or her income increased but not as much as the increase in income. Thus saving, defined as income less consumption, should increase when income grows and the aggregate saving rate, defined as the national saving–income ratio, should increase with aggregate income. According to this part of the *General Theory*, rich people saved, poor people did not; rich countries saved, poor countries did not. Despite his acceptance of this simple Keynesian formulation in 1944, Modigliani reports that he was not convinced that the saving–income ratio should rise with aggregate income, and began to systematically reconsider the Keynesian law in 1946. He was particularly unhappy with the notion that saving should be regarded as a luxury good that would be 'purchased' in greater quantities by the rich than poor in order to 'bequeath a fortune'. 'This explanation satisfied me not a jot' (2001, p. 52).

In the late 1940s Modigliani's alternative suggestion was that the saving–income ratio should fluctuate around a constant (or slowly moving) trend and that these fluctuations would be driven by the relationship of actual income to the normal income that the household could expect. In other

words, the household's saving rate was explained not by its absolute level of income (as Keynes would have it) but by its income *relative* to the aggregate mean income in the economy. Modigliani formulated his hypothesis in an elegant linear model in which the saving–income ratio was related negatively to the ratio of income at its previous peak to the current level of income (1949). When the economy was in recession (and current income was below its previous peak), saving and the saving–income ratio would both fall. This movement reflected the cyclical movement of consumption emphasized by Keynes. But, when the economy was growing and incomes were pushed above their previous peak, saving would rise and the saving–income ratio would return to its previous level. Thus the aggregate consumption function would shift upward in a ratcheting movement as aggregate income set new records.

Modigliani tested this formulation and estimated the parameters of the model using aggregate data for 1921 to 1940. James Duesenberry independently hit upon a very similar formulation. Duesenberry's 'relative income hypothesis' reconciled the time series and cross-sectional data by suggesting that the higher consumption of the poor was an attempt to keep up with those better situated economically. Both contributions were published in 1949. The differences in their theoretical justifications were generally glossed over by subsequent commentators and the empirical model became known as the Duesenberry–Modigliani hypothesis.

The success of the Duesenberry-Modigliani empirical work (and the growing sophistication of econometric technique) produced a flurry of follow-up empirical studies. Modigliani and his collaborator Richard Brumberg described the state of affairs, in a passage that reflects Modigliani's scientific philosophy (2001, p. 129). Empirical work should test theory; theory should be inspired by empirical observation; and progress would be made only through the constant interplay between the two:

> It may be said that, at the date of this writing (1952), the analysis of the consumption function has degenerated into a morass of seemingly contradictory, or at least disconnected, results, with each new empirical finding adding less to our understanding than to the existing confusion. Further empirical analysis is not likely to advance us very far until the economic theorist has been able to provide a conceptual framework to give coherence to past findings and guidance for the collection of more 'facts.'

Shortly after arriving at the University of Illinois, Modigliani began working with Brumberg to provide the missing conceptual foundation for the macroeconomic theory of consumption based on microeconomic marginal utility analysis. They produced two papers in 1952. The first, 'Utility Analysis and the Consumption Function,' was published in 1954. The other, 'Utility Analysis and Aggregate Consumption Functions,' was unpublished at the time of Blumberg's sudden and tragic death from a cerebral embolism in 1955. Modigliani was devastated by his friend's death and 'lost all interest in revising the manuscript' for publication (2001, p. 66). It remained unpublished for a quarter of century. It finally appeared in Modigliani's *Collected Papers* (1980) exactly as it had been left at the time of Blumberg's death. Together the two papers describe the life-cycle hypothesis (LCH).

The microeconomic model of consumption and saving proposed by Modigliani and Brumberg took the perspective of a forward-looking individual (or a couple) with a finite lifespan and no desire to bequeath a fortune to heirs (Keynes's proposed motive for aggregate saving was thus explicitly rejected). The model recognized that income will vary over the lifetime, rising at first as the individual's career advances and he or she gains experience and skill, but income will ultimately fall with age and may even disappear during retirement. With this view saving behaviour would vary over a person's lifetime. When young, the individual would save very little (when income is low relative to what can be expected in middle age). During the period of peak earnings in middle age, the individual's saving will be high as assets are accumulated to finance late life consumption and to afford retirement. When retired, the individual dissaves (saving is negative) as the accumulated assets are sold to support a planned retirement lifestyle.

The most familiar (and most simplified) exposition of the microeconomic model is that published by Modigliani in *Social Research* in 1966. That version pictured the expected income profile as flat and constant until retirement when it fell to zero and the desired consumption profile as flat throughout life. Over the lifespan the total of consumption would exactly exhaust the total income earned, but, since consumption must be maintained during the retirement years, consumption is less than income during the earning years. The 1966 article first introduced the diagram of the 'Modigliani pyramid', made famous by macroeconomic textbooks, that was reproduced in Modigliani's Nobel lecture, and which he came to view as his 'trademark' (2001, p. 60). In this diagram, the lifetime profile of wealth rises linearly with age until it reaches a maximum on the day of retirement and then declines linearly with age until death, thus tracing out the pyramid shape. In the more general case, the wealth profile would be hump-shaped.

In the elementary formulation of the model it was assumed for convenience (and also to make a sharp contrast to the common view) that individuals had no desire to make a net bequest to heirs; that is, they had no reason to make net accumulations in order to bequeath a greater inheritance then they had received. Modigliani had always argued, however, that a bequest motive could be added to the LCH without disturbing its implications. Yet he maintained that empirically a bequest motive would be 'relevant for the very rich (and especially for the *nouveaux riches*)'. At the same time Modigliani argued that in the absence of bequest motives there would still be substantial bequests left at death. If individuals knew the date of their death in advance, as assumed in the simplified exposition, then each individual over his lifetime would consume one hundred per cent of his or her lifetime income. The 'life-time propensity to consume' would be 1. Since people do not generally foresee the timing of their death, they must plan their saving to be sufficient to support them to a very old age. Since, alas, many die at a younger age then this, inheritance bequests are commonplace, but are for the most part unintended.

The importance of Modigliani–Brumberg *micro*economic model of saving lies in the *macro*economic implications of life-cycle behaviour. Aggregate saving in the LCH does not depend upon current income but on life-cycle income. Thus the age structure of the population matters. In a population that is growing rapidly because of natural increase or immigration, there will be more young and middle-aged savers than older retired dissavers. Aggregate saving will be higher. Likewise, in an economy that is experiencing rapid economic growth, perhaps produced by new technologies and strong investment in new capital, the young and middle-aged savers will look forward to higher lifetime earnings while the older dissavers are consuming at a level commensurate with their assets accumulated over a lifetime when productivity was lower. Thus growth is good for saving. Moreover, the higher aggregate rates of saving generated by either population growth or by economic growth can help sustain the forward progress by financing investment at continuing high levels. Saving is good for growth.

Another important long-run implication of the life-cycle hypothesis is that sustained government deficits will be a drag on economic growth. Modigliani called attention to the burden of the national debt in a famous paper published in the *Economic Journal* (1961). The government finances its deficit spending by issuing government bonds which are a form of net worth for those who purchase them. When members of the public hold some of their wealth in the form of bonds, the bonds substitute for the physical capital (machines, structures, and other productive capital) that would otherwise be created to satisfy the demand for life-cycle assets. The burden of the national debt is the reduced rate of growth attributable to the reduced rate of capital formation. This burden can be said to fall on future generations by reducing their income below what it would be otherwise.

Modigliani's analysis of the burden of the national debt was criticized by Robert Barro (1974). Barro's approach to the issue is also known as the 'Ricardian equivalence theorem' because it echoed a suggestion of David Ricardo. Barro rejected Modigliani's view that an

individual's planning horizon is constrained to his expected lifetime and took the extreme opposite position that the planning horizon is infinite. Government deficits today, Barro argued, should lead to an increase in saving as taxpayers reasoned that taxes would have to be raised in some indefinite future to pay off the debt. To have the assets needed to meet this forecast tax increase, taxpayers would temporarily increase saving to set the required sum aside. Modigliani viewed Barro's assumption of an infinite horizon as 'incredible' and the equivalence theorem 'untenable' (2000, p. 235). In characteristic fashion, however, he responded with carefully designed empirical tests rather than theoretical debate. In his presentation of the data, the LCH and the burden of the debt were supported and Ricardian equivalence rejected (1983a, 1986b).

The simple version of the LCH made no allowance for the Social Security pension system as an alternative to private saving. Modigliani argued that incorporating a mandatory government retirement plan into the model was straightforward and, more importantly, that treating Social Security consistently would clear up several important misunderstandings. As he would model it, Social Security's payroll taxes should be considered a form of forced or 'compulsory' saving that builds up 'Social Security wealth'. The benefits received in old age should then be seen as drawing down those assets. When Social Security is included as a form of wealth, the empirical wealth profile has the hump shape predicted by the LCH (1983 with Arlie Sterling; 1987 and 2005 with Tullio Jappelli). This answered those critics who failed to find much dissaving in old age when using a conventional definition of saving. The critics had simply defined wealth too narrowly.

The introduction of Social Security into an economy that previously relied exclusively on private saving, according to Modigliani, would have two effects on the private saving rate. One is the replacement effect. Because the Social Security tax is a form of forced saving, individuals who count on the promised benefits can save less and on this account the society's wealth–income ratio will be reduced. On the other hand, there might be an offsetting 'retirement effect'. A Social Security system will encourage earlier retirement both directly and through a social emulation effect. Longer retirement periods require greater wealth accumulation and thus increased saving rates. Empirical work reported by Modigliani and his coauthor Arlie Sterling suggests that the two effects roughly cancel each other out (1983).

The long-run implications of the LCH that saving is increased by economic growth, that the national debt produces a burden, and that there is little reason to think that the introduction of Social Security significantly reduced the saving rate challenged conventional views at the time. Not surprisingly, there were many critics. Modigliani's persistent defence of the logic of the theory and his continuous production (with the help of many coauthors) of ingeniously designed and carefully executed empirical verifications and rejoinders kept the model in the forefront of academic analysis and policy debate. It remains the accepted view.

Yet it was the short-run or cyclical implications – not the long-run consequences – that received the more immediate attention. In a pair of papers coauthored with Albert Ando, Modigliani directed attention to the short-run considerations and the implications for the aggregate time-series consumption function (1963, 1965). The underlying theory had been formulated in the still unpublished second paper with Richard Brumberg but the work with Ando brought the cyclical implications to the attention of the profession. The short-term consumption function proposed by Ando and Modigliani made consumption a linear function of aggregate disposable *labour* income (that is, income excluding the return to asset holdings and less the amount of personal taxes) and aggregate net worth. The coefficients of the two variables could be taken as empirically constant in the short run determined by the length of life, the length of retirement, and the rate of growth. It was not until estimates of the aggregate stock of net worth became available that the model could be verified empirically. When Raymond Goldsmith published his wealth estimates (1962), the life-cycle consumption function passed the battery of tests designed by

Ando and Modigliani with the highest marks (Ando and Modigliani 1963; Modigliani 1966).

The cyclical properties of the LCH equation were not in themselves particularly novel. The LCH behaved in the short run not unlike the Duesenberry–Modigliani model or the roughly contemporaneous theory of consumption put forward by Milton Friedman, the permanent income hypothesis (1957). The saving–income ratio would fall during recessions and rise during upturns, but would fluctuate about a fairly stable long-run average. And, like the simpler Keynesian model, the Ando–Modigliani formulation implied that tax cuts could stimulate consumption and thus help counteract recessionary tendencies. There were, however two novel implications of the cyclical formulation of the LCH with important policy implications. The short-run life-cycle consumption function postulates that consumption would be responsive to the value of assets; thus a stock market crash, like that of 1929, would tend to reduce consumption as individuals sought to restore their lost wealth. This was not an implication of the alternative models. As another contrast, Friedman suggested that consumption each period should depend upon the current rate of interest since consumers would be willing to save more (and consume less) when the reward to asset holding is high. Modigliani conjectured that the saving rate would be 'largely independent' of the interest rate. While accepting Friedman's point that an increase in the reward for saving (higher interest rates) would induce an increase in saving, Modigliani pointed out another consequence of high rates. High interest rates would allow the stock of assets to accumulate more rapidly, thus requiring less saving to reach the target level of assets needed for retirement. Modigliani suggested the two effects would largely cancel out. If Modigliani is correct, short-term policy strategies to increase saving by manipulating the rate of interest would be ruled out.

Expectations and Fluctuations

One of Keynes's foremost contributions, according to his own view, was his emphasis on the importance of expectations. The central conclusion of his *General Theory*, announced in the Preface, was that a 'monetary economy ... is essentially one in which changing views about the future are capable of influencing the quantity of employment and not merely its direction' (1936). It is somewhat ironic, then, that Modigliani's 1944 reformulation of the *General Theory* took expectations as given. We are told that this was a simplification for 'convenience' since the paper was concerned with 'the determinants of equilibrium, and not with the explanation of business cycles' (1944, p. 46). Most of the equations of his equilibrium model, to be sure, contain variables that represent the expectations of economic agents. But, to take account of any relevant *change* of views about the future, the analyst would have to shift one or more of the relationships expressed in the system of equations.

A few years after formulating the equilibrium model with static expectations, Modigliani began a far-ranging investigation of the role of anticipations and uncertainty in the explanation of business cycles. In 1949 Modigliani began work on a project he called 'Economic Expectations and Fluctuations'. It would occupy him for more than ten years. Modigliani moved the project to Carnegie Tech in 1952. There he collaborated with Herbert Simon, Charles Holt, and John Muth (1960) and Kalman Cohen (1961) on two books concerned with anticipations, forecasting, and the use of inventories to smooth production. While the work on the life cycle and production smoothing explicitly recognized the importance of expectations in microeconomic models, a breakthrough came when Modigliani turned his attention to modelling the formation of expectations in a macroeconomic context. Modigliani collaborated with Emile Grunberg, a colleague at Carnegie, on the 'Predictability of Social Events' (1954), a famous paper that is widely recognized as introducing the concept of 'rational expectations' into economic theory. The concept itself is simple. Rational expectations are forecasts of the future that are consistent with the way the economy is believed to work. To adopt any other expectation would be to ignore whatever knowledge one had about the workings of the economy.

The macroeconomic implications of rational expectations, however, proved to be profound. In the hands of others, rational expectation formation was used to question the effectiveness of Keynesian monetary and fiscal stabilization policy, and thus Modigliani's 'invention' reappeared later as a challenge to the legitimacy of Keynesian economics.

The problem that Grunberg and Modigliani set out to explore was whether a widely believed public prediction of a future event might change individuals' behaviour in such a way as to invalidate the prediction. Their answer was that a correct private prediction would be a wrong public prediction. Nevertheless, accurate public prediction was possible because the reaction of the public to the announcement can be taken into account by the social scientist. Accurate public predictions are predictions that are 'internally consistent' in the sense that they recognize and incorporate any change in public expectations induced by the prediction itself that would influence the course of events.

It was left to Modigliani's student at Carnegie, John Muth, to extend the concept of internally consistent expectations to become 'rational expectations', an exercise Muth (1961) carried out in a microeconomic context. Ten years later Robert Lucas (1972) returned the concept to a macroeconomic setting (in the context of a market-clearing model) and suggested that stabilization policies could not change real output in a predictable way if those policies were fully anticipated. The macroeconomic rational expectations model became the foundation of the 'new classical macroeconomics', so called because money had no real effects in this model. These developments took place in the 1970s and were led by others; meanwhile, Modigliani's thinking about expectations had been developing in another direction during the 1950s and 1960s. In his paper with Brumberg, Modigliani argued that, while anticipations about the future life course would be relevant to the individual's decision about current consumption, it was not necessary to take explicit account of uncertainty about the future. Uncertainty would simply give rise to an additional precautionary motive for saving, but the assets accumulated to satisfy the life-cycle motive would do double duty as a buffer stock to insure against emergencies. In making this argument Modigliani was echoing arguments that he had advanced in the books on business planning. In this work Modigliani observed firm behaviour and offered a description of how businesses form expectations about the future and how they make use of those anticipations in current decision making. The upshot was that, while knowledge about the future would be important, firms need not (and therefore did not) attempt to acquire all possible information. Much information about the future would be beyond the relevant planning horizon and therefore irrelevant, other information would not need to be precise, and some information might not be worth the effort to acquire. Businesses in the real world attempt to make the best possible forecasts of the variables deemed important, but sometimes best practice would be a rule of thumb or a simple extrapolation of the past. Any uncertainty that remained would be adequately hedged since inventory could do double duty and serve as a buffer stock against inadequately foreseen contingencies as well as smooth production.

This is a clearly pragmatic approach to expectations. Modigliani suggested that a pragmatic formulation was realistic. The 'expectation function will, at best, appear in the form of broad statistical generalizations' since expectations about the future range in practice 'from the elaborate scientific forecast of the large business enterprise to primitive guesses and dark hunches' (Grunberg and Modigliani 1954, p. 471). It was this realistic approach to expectations that Modigliani later carried over to macroeconomic models. It is important to note, however, that Modigliani was not opposed to the idea of rational expectations in principle. He declared the concept 'a good starting point' for analysis and thought the assumption would be 'sensible' in some circumstances, for example, in financial markets (1983b, pp. 123–4). He considered Muth's contribution 'fundamental' and an improvement over 'naïve or *ad hoc* assumptions' regarding the formation

of expectations (1986b, p. 25). But when rational expectations were used to support the new classical economics and its startling proposition that stabilization policy would be ineffective, he thought that this was 'pushing the idea of rationality well beyond the range where it is useful' (1983b, p. 123). 'It is a 'wonderful theory ... [but] it is *not* a description of the world' (2000, p. 235).

Characteristically, Modigliani was not content to simply debate the logical merits of a model or the realism of its assumptions. The new classical macroeconomics could be rejected because its conclusions were inconsistent with the empirical evidence. The model implied that fluctuations in unemployment should be mild, short-lived, and random, contrary to all experience. The new classical model was also inconsistent with the existence of long-term contracts. If such contracts are rational, then wages are rigid, contrary to a postulate of the new classical view. If they are not rational, then they 'should have long ago disappeared'.

Modigliani's pragmatic approach to modelling expectations was put to work in a series of papers on the term structure of interest rates (Modigliani and Sutch 1966, 1967; Modigliani and Shiller 1973). The 'term structure' refers to the relationship between interest rates on assets with different terms to maturity. In his collaboration with Sutch the long-term rate of interest was linked to the short-term rate through financial arbitrage. Since the investor could obtain a return over the long term by investing either in a long-term bond or alternatively in a sequence of short-term bills, the choice between the two would be influenced by the investor's expectations about the future course of short-term interest rates. Because the expectations would be subject to uncertainty, each investor would have a natural preference for assets with a maturity that matched their needs. But they could be tempted out of this 'preferred maturity habitat' if the advantage with shorter or longer maturities were forecast to be large enough.

An empirical characterization of expectation formation was required to complete the model of the relationship between short- and long-term rates. Here again Modigliani looked to how investors actually behaved. Modigliani and Sutch suggested that future expected rates were formulated by extrapolating past movements. They proposed that the recent trend in the rate would be anticipated to continue for a while, but that the best guess for the long run was that rates would return to their long historical average (as Keynes had suggested). Modigliani and Sutch considered this formulation a 'plausible' representation of how investors actually thought about the problem. Modigliani and Shiller went on to demonstrate that the Modigliani–Sutch model of expectations was also rational in the sense that it represented the best forecast possible on the basis of all information available.

Corporate Finance

A key component of the Keynesian macroeconomic structure is the investment function, which held that the aggregate volume of investment would be responsive to the cost of capital. In the 1950s Modigliani turned his attention to this topic as well. The result was spectacular. In citing Franco Modigliani for the Nobel Prize in 1986, the Nobel Foundation's judges singled out both the life-cycle hypothesis and the path-breaking Modigliani–Miller theorems on corporate dividends, leverage, and the cost of capital (Modigliani and Miller 1958, 1963; Miller and Modigliani 1961; Modigliani 1982). The two MM theorems, as they are called, not only overturned the existing thinking about the cost of capital but launched modern finance theory. Indeed, this line of research was deemed so important that Merton Miller later received his own Nobel Prize in 1990 for his contribution to the joint work. In 1956 Merton Miller was an assistant professor auditing Modigliani's course at the Carnegie Institute of Technology. He became excited when Modigliani introduced the topic in class and agreed to join him in working out the proof.

The first Modigliani–Miller theorem establishes, when a firm's investment policy is fixed, that the market evaluation of a firm would be unaffected by its volume of debt in a simplified world with well-functioning financial markets,

rational investors, and neutral taxes (1958). The second theorem, an extension of the first, states under the same assumptions that the value of the firm is independent of its dividend policy (1961). Taken together they suggest that 'financial policy does not matter!' (1982, p. 255).

The contribution to the scientific analysis of finance was profound. First, the MM papers introduced the application of microeconomic theory – and in particular the notion of arbitrage – to problems in corporate finance. Rigorous mathematical modelling has been the hallmark of the field ever since. Second, the two theorems taken together allow the separation for analytical and management purposes of investment decisions from financial decisions. The implication for the structuring of corporate management has led over time to the division of managerial responsibilities between the CEO and the CFO. Third, the MM theorems were established in the context of a highly stylized model, so a good deal of subsequent theoretical and empirical work has been devoted to understanding the impact of relaxing the simplifying assumptions and extending the application of the model. This research agenda has enriched the field immeasurably. The MM theorems, for example, have led directly to subsequent developments in the evaluations of options.

At the time the first MM theorem was published it was held to be self-evident that borrowing and taking on debt would lower the cost of capital to the firm because the rate of interest on the loans was below the cost of raising capital through the sale of equity. Modigliani and Miller elegantly demonstrated that the old theory was seriously flawed. As is typical of Modigliani's work, the result rests on the clear application of a microeconomic principle, in this case, the role of arbitrage. The intuition behind the two theorems is simple. No matter what the debt–equity structure of the firm (or its dividend–retained earnings policy) the investor can always undo the impact on his or her own stock portfolio by adding or subtracting other equities or forms of debt to the mix. The resulting arbitrage will mean that the market value of the firm will depend only on the income stream generated by its assets.

Despite the enormous literature that the MM theorems generated and despite the transformation of the field of corporate finance as a consequence, Modigliani was fond of trivializing the idea behind MM as 'obvious' and said that the theorems were written with 'tongue-in-cheek' as a way of chastising the 'old school' of finance for its reliance on anecdotes and rules of thumb relayed through case studies and the reminiscences of managers and accountants (2000, pp. 233–4). Yet the papers from 1958 and 1963 are among the top three most-cited papers by Modigliani and also among those he listed as his 'personal favorites' (Merton 1987).

Other Topics

In 1997 Modigliani published with Leah Modigliani, his granddaughter and a financial analyst, a paper entitled 'Risk-Adjusted Performance: How to Measure It and Why'. Together they proposed a measure of the rate of return on an investment portfolio that was adjusted for risk so that the performance of different fund portfolios could be compared with the same measuring rod. Their technique of risk adjustment is now widely used on Wall Street and has become known as M^2 – 'M-squared' – for the two Modiglianis. It applies the same concept of arbitrage introduced by Modigliani and Miller to neutralize the risk. If the historical volatility of a portfolio has been high relative to a benchmark (say the S&P 500), its risk could hypothetically be reduced to match that of the benchmark by adding treasury bills in sufficient quantity to the mix. If the investment portfolio under study has a volatility below that of the benchmark, it could hypothetically be levered up to match the risk standard by borrowing on margin and investing additional sums in the fund. The rates of return can then be calculated and compared for these blended portfolios.

A longer review of Modigliani's work would have to find space to discuss his writings on the Italian economy with La Malfa (1967), Tarantelli

(1975), Padoa-Schioppa and Rossi (1986), and Jappelli (1987) and on the European economy and international finance (Part III of *Collected Papers*, vol. 3), to mention just a few of the papers that were published in English. For Modigliani's recounting of this work, much of which is in Italian, see chapters "▶ Millar, John (1735–1801)" and "▶ Collet, Clara Elizabeth (1860–1948)" of his autobiography, *Adventures of an Economist* (2001). A review of those contributions will suggest that there are serious omissions from the present survey.

Retrospect

Modigliani was a brilliant economist who took the real problems of the real world seriously and then developed powerful theories to explain what he saw. Yet he was uncomfortable with his theories until he had rigorously tested them against the data and against alternative explanations. These econometric explorations invariably stimulated him to re-examine his thinking. For him, it was a process without end. Modigliani rarely let a topic go, he worked continuously to refine, improve, and, when necessary, to defend each of his signature contributions. He was driven by a strong faith in the power of an economic theory so derived to inform policy, to solve problems, and to right social wrongs. He acted on his beliefs by becoming an advisor to – or, when they would not listen, a public opponent of – those who made economic policy. He changed his mind when logic or facts dictated it, yet he remained steadfast in his belief that economics was a science with the potential to make the world a better place. This combination of dedication, intellectual honesty, and liberal values made him one of the most influential macroeconomists of the 20th century.

His personal characteristics were an important ingredient of his success. He was warm and caring, intense and excitable, enthusiastic and full of an infectious joy for life. He was a charismatic teacher, a tenacious debater, and a seminal thinker with the rare ability to stimulate others to think and imagine beyond their usual capacity.

See Also

- ▶ Expectations
- ▶ Keynesian Revolution
- ▶ Miller, Merton (1923–2000)
- ▶ Modigliani–Miller Theorem
- ▶ Rational Expectations
- ▶ Term Structure of Interest Rates

Selected Works

The most important scientific contributions of Modigliani have been collected and reprinted in *The collected papers of Franco Modigliani*, 6 vols., ed. A. Abel, S. Johnson and F. Franco (1980–2005), Cambridge: MIT Press.

1944. Liquidity preference and the theory of interest and money. *Econometrica* 12, 45–88.

1949. Fluctuations in the saving–income ratio: A problem in economic forecasting. In *Studies in income and wealth*, vol. 11. New York: NBER.

1954. (With R. Brumberg.) Utility analysis and the consumption function: An interpretation of cross-section Data. In *Post-Keynesian economics*, ed. K. Kurihara. New Brunswick: Rutgers University Press.

1954. (With E. Grunberg.) The predictability of social events. *Journal of Political Economy* 62, 465–478.

1958. (With M. Miller.) The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48, 261–297.

1960. (With C. Holt, J. Muth and H Simon.) *Planning production, inventories, and work force*. Englewood Cliffs: Prentice-Hall.

1961. Long-run implications of alternative fiscal policies and the burden of the national debt. *Economic Journal* 71, 730–755.

1961. (With K. J. Cohen.) *The role of anticipations and plans in economic behavior and their use in economic analysis and forecasting*. Urbana: University of Illinois.

1961. (With M. Miller.) Dividend policy, growth, and the valuation of shares. *Journal of Business* 34, 411–433.

1963. The monetary mechanism and its interaction with real phenomena. *Review of Economics and Statistics* 45(1) Part 2, Supplement, 79–107.
1963. (With A. Ando.) The 'life-cycle' hypothesis of saving: Aggregate implications and tests. *American Economic Review* 53, 55–84.
1963. (With M. Miller.) Corporate income tax and the cost of capital: A correction. *American Economic Review* 53, 433–443.
1964. Some empirical tests of monetary management and of rules versus discretion. *Journal of Political Economy* 72, 211–245.
1965. (With A. Ando.) The relative stability of monetary velocity and the investment multiplier. *American Economic Review* 55, 693–728.
1966. The life-cycle hypothesis of saving: The demand for wealth and the supply of capital. *Social Research* 33, 160–217.
1966. (With R. Sutch.) Innovations in interest rate policy. *American Economic Review* 56, 178–197.
1967. (With G. La Malfa.) Inflation, balance of payments deficit, and their cure through monetary policy: The Italian example. *Banca Nazionale del Lavoro Quarterly Review* 80, 3–47.
1967. (With R. Sutch.) Debt management and the term structure of interest rates: An empirical analysis. *Journal of Political Economy* 75, 569–589.
1969. (With A. Ando.) Econometric analysis of stabilization policies. *American Economic Review* 59, 296–314.
1973. (With R. Shiller.) Inflation, rational expectations and the term structure of interest rates. *Economica* 40, 12–43.
1975a. Channels of monetary policy in the Federal Reserve-MIT-University of Pennsylvania Econometric Model of the United States. In *Modelling the economy*, ed. G. Renton. London: Heinemann Educational Books.
1975b. The life-cycle hypothesis of saving twenty years later. In *Contemporary issues in economics*, ed. M. Parkin and A. Nobay. Manchester: Manchester University Press.
1975. (With E. Tarantelli.) The consumption function in a developing economy and the Italian experience. *American Economic Review* 65, 825–842.
1977. The monetarist controversy or, Should we forsake stabilization policies? Presidential address delivered at the American Economic Association. *American Economic Review* 67, 1–19.
1980. (With R. Brumberg.) Utility analysis and aggregate consumption functions: An attempt at integration. In *The collected papers of Franco Modigliani*, vol. 2, ed. A. Abel. Cambridge, MA: MIT Press.
1982. Debt, dividend policy, taxes, inflation, and market valuation. Presidential address delivered at the American Finance Association. *Journal of Finance* 37, 255–273.
1983a. Government deficits, inflation, and future generations. In *Deficits: How big and how bad?*, ed. D. Conklin and T. Courchene. Ontario: Ontario Economic Council.
1983b. Interview with Franco Modigliani. In *Conversations with economists: New classical economists and their opponents speak out on the current controversy in macroeconomics*, ed. A. Klamer. Totowa: Rowman and Littlefield.
1983. (With A. Sterling.) Determinants of private saving with special reference to the role of Social Security – cross-country tests. In *The determinants of national saving and wealth*, ed. F. Modigliani and R. Hemming. London: Macmillan.
1986a. Autobiography. *Les Prix Nobel. The Nobel Prizes 1985*, ed. W. Odelberg. Stockholm: Nobel Foundation. Online. Available at: http://nobelprize.org/nobel_prizes/economics/laureates/1985/modigliani-autobio.html. Accessed 9 Sept 2006.
1986b. Life cycle, individual thrift, and the wealth of nations. Nobel Prize Lecture, 1985. Reprinted in *American Economic Review* 76, 297–313.
1986c. *The debate over stabilization policy*. Cambridge: Cambridge University Press.
1986. (With F. Padoa-Schioppa and N. Rossi.) Aggregate unemployment in Italy, 1960–1983. *Economica* 53, S245–S273, S347–S352.
1986. (With A. Sterling.) Government debt, government spending, and private sector behavior: Comment. *American Economic Review* 76, 1168–1179.
1987. (With T. Jappelli.) Fiscal policy and saving in Italy since 1860. In *Private saving and*

public debt, ed. M. Boskin, J. Flemming and S. Gorini. Oxford: Basil Blackwell.
1988a. The role of intergenerational transfers and life cycle saving in the accumulation of wealth. *Journal of Economic Perspectives* 2(2), 15–40.
1988b. MM – past, present, future. *Journal of Economic Perspectives* 2(4), 149–58.
1997. (With L. Modigliani.) Risk-adjusted performance: How to measure it and why. *Journal of Portfolio Management* 23(2), 45–54.
2000. An interview with Franco Modigliani. Interviewed by W. Barnett and R. Solow, 5–6 November 1999. *Macroeconomic Dynamics* 4, 222–256.
2001. *Adventures of an Economist*. New York: Texere.
2005. (With T. Jappelli.) The age–saving profile and the life–cycle hypothesis. In *The collected papers of Franco Modigliani*, vol. 6, ed. F. Franco. Cambridge: MIT Press.

Bibliography

Barro, R. 1974. Are government bonds net worth? *Journal of Political Economy* 82: 1095–1117.
Duesenberry, J. 1949. *Income, saving, and the theory of consumer behavior*. Cambridge: Harvard University Press.
Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.
Goldsmith, R. 1962. *The national wealth of the United States in the postwar period*. Princeton: Princeton University Press.
Hicks, J. 1937. Mr Keynes and the 'Classics'. *Econometrica* 5: 147–159.
Keynes, J.M. 1936. *The general theory of employment, interest, and money*. New York: Harcourt, Brace.
Lucas, R. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Merton, R. 1987. In honor of Nobel Laureate, Franco Modigliani. *Journal of Economic Perspectives* 1(2): 145–155.
MIT (Massachusetts Institute of Technology). 2003. Nobel laureate Franco Modigliani dies at 85. News Office, MIT. Online. Available at: http://web.mit.edu/newsoffice/2003/modigliani.html. Accessed 11 Nov 2003.
Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.

Partial support for this research was provided by the National Science Foundation and the Center for Social and Economic Policy at the University of California, Riverside.

Modigliani–Miller Theorem

Anne P. Villamil

Abstract

The Modigliani–Miller theorem provides conditions under which a firm's financial decisions do not affect its value. The theorem is one of the first formal uses of a no arbitrage argument and it focused the debate about firm capital structure around the theorem's assumptions, which set the conditions for effective arbitrage. The search for the source of the 'failure of irrelevance' has led to important advances in the nature of financial structure, and more fundamentally to the types of frictions that would cause agents to have different market opportunities, information sets or commitment frictions.

Keywords

Agency problems; Arbitrage; Bankruptcy; Capital gains; Control rights; Debt versus equity; Debt–equity ratio; Dividend policy; Finance; Imperfect information; Incomplete contacts; Information revelation; Insider trading; Law of one price; Leasing vs. buying; Miller, M.; Modigliani, F.; Modigliani–Miller theorem; Monitoring; Optimal capital structure; Optimal contracts; Separating equilibrium; Separation of ownership and control; Shareholder voting rights; Signalling models of equity; Taxation of capital income; Verification costs

JEL Classifications

G3

The Modigliani–Miller theorem is a cornerstone of modern corporate finance. At its heart, the theorem is an irrelevance proposition: it provides conditions under which a firm's financial decisions do not affect its value. Modigliani explains the theorem as follows:

> ... with well-functioning markets (and neutral taxes) and rational investors, who can 'undo' the corporate financial structure by holding positive or negative amounts of debt, the market value of the firm – debt plus equity – depends *only* on the income stream generated by its assets. It follows, in particular, that the value of the firm should not be affected by the share of debt in its financial structure or by what will be done with the returns – paid out as dividends or reinvested (profitably). (Modigliani 1980, p. xiii)

In fact, what is currently understood as the Modigliani–Miller theorem comprises four distinct results from a series of papers (1958; 1961; 1963). The first proposition establishes that under certain conditions, a firm's debt–equity ratio does not affect its market value. The second proposition establishes that a firm's leverage has no effect on its weighted average cost of capital (that is, the cost of equity capital is a linear function of the debt–equity ratio). The third proposition establishes that firm market value is independent of its dividend policy. The fourth proposition establishes that equity-holders are indifferent about the firm's financial policy.

Miller (1991, p. 5) explains the intuition for the theorem with a simple analogy. 'Think of the firm as a gigantic tub of whole milk. The farmer can sell the whole milk as it is. Or he can separate out the cream, and sell it at a considerably higher price than the whole milk would bring.' He continues: 'The Modigliani–Miller proposition says that if there were no costs of separation (and, of course, no government dairy support programme), the cream plus the skimmed milk would bring the same price as the whole milk.' The essence of the argument is that increasing the amount of debt (cream) lowers the value of outstanding equity (skimmed milk) – selling off safe cash flows to debt-holders leaves the firm with more lower-valued equity, keeping the total value of the firm unchanged. Put differently, any gain from using more of what might seem to be cheaper debt is offset by the higher cost of now riskier equity. Hence, given a fixed amount of total capital, the allocation of capital between debt and equity is irrelevant because the weighted average of the two costs of capital to the firm is the same for all possible combinations of the two.

The theorem makes two fundamental contributions. In the context of the modern theory of finance, it represents one of the first formal uses of a no arbitrage argument (though the 'law of one price' is long-standing). More fundamentally, it structured the debate on why irrelevance fails around the theorem's assumptions: (i) neutral taxes; (ii) no capital market frictions (that is, no transaction costs, asset trade restrictions or bankruptcy costs); (iii) symmetric access to credit markets (that is, firms and investors can borrow or lend at the same rate); and (iv) firm financial policy reveals no information. Modigliani and Miller (1958) also assumed that each firm belonged to a 'risk class', a set of firms with common earnings across states of the world, but Stiglitz (1969) showed that this assumption is not essential. The relevant assumptions are important because they set conditions for effective arbitrage: When a financial market is not distorted by taxes, transaction or bankruptcy costs, imperfect information or any other friction which limits access to credit, then investors can costlessly replicate a firm's financial actions. This gives investors the ability to 'undo' firm decisions, if they so desire. Attempts to overturn the theorem's controversial irrelevance result were a fortiori arguments about which of the assumptions to reject or amend. The systematic analysis of these assumptions led to an expansion of the frontiers of economics and finance.

The importance of taxes for the irrelevance of debt versus equity in the firm's capital structure was considered in Modigliani and Miller's original paper (1958). Modigliani and Miller (1963) and Miller (1977) addressed the issue more specifically, showing that under some conditions, the optimal capital structure can be complete debt finance due to the preferential treatment of debt relative to equity in a tax code. For example, in the United States, interest payments on debt are excluded from corporate taxes. As a consequence, substituting debt for equity generates a surplus by reducing firm tax payments to the government. Firms can then pass this surplus on to investors in the form of higher returns. This raised the further provocative question – were firms that issued equity leaving stockholder money on the table in the form of unnecessary corporate income

tax payments? Miller (1977) resolved this problem by showing that a firm could generate higher after-tax income by increasing the debt–equity ratio, and this additional income would result in a higher payout to stockholders and bondholders, but the value of the firm need not increase. The crux of the argument is that as debt is substituted for equity, the proportion of firm payouts in the form of interest on debt rises relative to payouts in the form of dividends and capital gains on equity. Taxes that are higher on interest payments than on equity returns reduce or eliminate the advantage of debt finance to the firm.

The remaining Modigliani–Miller assumptions deal with various types of capital market frictions (for example, transaction costs or imperfect information) that are at the heart of arbitrage. The driving force in a perfect market for a homogeneous good is the 'law of one price'. If debt and equity are merely different packages of an underlying homogeneous good – capital – and there are no market imperfections, then it follows immediately that the law of one price holds due to arbitrage. Investors simply engage in arbitrage until any deviation in the price of the two forms of capital is eliminated. Thus, the remaining discussion is organized around the implications of the theorem for firm capital structure, dividend policy, and the method of capital finance (lease versus buy).

With regard to firm capital structure, the theorem opened a literature on the fundamental nature of debt versus equity. Are debt and equity distinct forms of capital? Why and in what specific ways? In order to answer these questions about the nature of capital, the optimal contract literature examines debt and equity as financial contracts that arise optimally in response to particular market frictions, when contracting possibilities are complete or incomplete. Complete contracts can be written on all states if this is optimal; incomplete contacts cannot depend on some states of nature.

In one of the earliest contributions, Townsend (1979) combines elements of imperfect information and bankruptcy costs to examine the nature of debt in a complete contracting environment. In his costly state verification model, debt is an optimal response to costly monitoring and differential information: all agents know *ex ante* the distribution of firm returns, but only the firm privately and costlessly observes the return *ex post*. The lender can acquire this information, but must irrevocably commit to pay a deadweight verification cost. Townsend shows that debt is optimal because it minimizes this cost. When the firm makes the required fixed debt repayment, no cost is incurred. Only when the firm is insolvent, and hence cannot repay its debt fully, does verification occur. Townsend interprets this as costly bankruptcy (liquidation): the firm is shut down; firm assets are seized by a 'court', which verifies their magnitude and transfers the residual to the lender, net of the verification cost. Lacker and Weinberg (1989) extend the approach by specifying conditions under which equity is optimal in an analogue of the model, costly state falsification. Neither debt nor equity is *ex post* efficient in this class of models because no agent wishes to request costly intervention and incur the deadweight cost, *ex post*. Agents know that bankruptcy occurs only when the firm is truly unable to repay due to a low realization, but they are implicitly assumed to be committed to the decisions they made *ex ante*. Otherwise, debt is no longer optimal.

Krasa and Villamil (2000) show that a firm–lender investment problem with multiple stages, costly enforcement, limited commitment and an explicit enforcement decision, can illuminate debt's distinct properties. The analysis also solves the *ex post* inefficiency problem in the costly state verification model. Agents write a contract in the initial period, knowing only the distribution of project returns. The contract specifies payments and when enforcement will occur, and can be altered if agents receive new information. In the next period, the borrower privately observes the return and can make the unenforceable payment specified in the original contract or propose an alternative payment (that is, renegotiate). In the final stage the investor can seek costly enforcement of the contractually specified payment or renegotiate enforcement. The opportunity to renegotiate is important because it introduces a new source of information: any positive renegotiation payment by the firm would reveal information to the investor about the firm's state. Debt is optimal because it minimizes information revelation. Renegotiation,

which imposes a constraint on the contract problem, is only relevant when an agent acquires new information and can use the information to alter the initial contract. Debt weakens agents' incentive to renegotiate by minimizing information revelation (a fixed face value reveals no information about the firm). The contract is *ex post* efficient because all decisions are chosen optimally as part of a perfect Bayesian Nash equilibrium. This minimal information revelation of debt stands in sharp contrast to the active information revelation in signalling models of equity. For example, in Leland and Pyle (1977) retained equity by a firm signals a profit increase sufficient to offset the owner's forgone diversification. In Myers and Majluf (1984), issuing equity signals bad news – owners with inside information sell shares when markets overvalue them. These signalling models leave open why a firm would use financial decisions to reveal information, a problem that does not arise in Krasa and Villamil.

In incomplete contracting models, control rights are an alternative justification for debt and equity contracts. Aghion and Bolton (1992) view debt as a particular assignment of control rights with important incentive properties. They show that when contracting possibilities are exogenously incomplete and control rights are assigned entirely to the investor or the firm, the first-best contract cannot be implemented. If the investor has sole control, the investor may force the firm to expand to a suboptimal level. Alternatively, if the firm has sole control it may not liquidate optimally. Aghion and Bolton show that, under some conditions, debt is the optimal contract because it assigns control to the firm in good states but to the investor in bad states. This ensures that optimal decisions are made in solvency and default states. Zender (1991) extends the model to include both debt and equity contracts. Grossman and Hart (1988) and Harris and Raviv (1988) examine control in the context of voting rights. They focus on the 'one vote per share' property of equity and majority voting, showing circumstances under which equity is optimal and when other 'extreme securities' are optimal.

Instead of focusing on the properties of debt and equity per se, Allen and Gale (1988, 1991) examine the properties of optimal securities more broadly, especially financial innovation. They study the problem of a firm that can issue securities in a market where the transaction cost of issuing securities makes the market incomplete. Market structure is endogenous in the sense that firms choose the securities they issue, which determines the transaction costs they incur. Allen and Gale (1988) prohibit short sales and show that neither debt nor equity is optimal. In contrast, Allen and Gale (1991) permit unlimited short sales, and show by example that debt and equity can be optimal. They note that the example is a special case; in general their model predicts that optimal securities are much more complex than those typically observed. The debt–equity puzzle unleashed by Modigliani and Miller continues to be an active area of research. The common theme of both the complete and incomplete contracting literatures is that debt, equity, and hybrid securities arise endogenously to overcome frictions in capital markets. Debt and equity have unique properties that resolve these frictions.

The Miller and Modigliani (1961) and Miller (1977) result that firm value is independent of dividend policy has also been examined extensively. Bhattacharya (1979) and others show that firm dividend policy can be a costly device to signal a firm's state, and hence relevant, in a class of models with: (i) asymmetric information about stochastic firm earnings; (ii) shareholder liquidity (a need to sell makes firm valuation relevant); and (iii) deadweight costs (to pay dividends, refinance cash flow shocks or cover underinvestment). In a separating equilibrium, only firms with high anticipated earnings pay high dividends, thus signalling their prospects to the stock market. As in other costly signalling models, the question as to why a firm would use financial decisions to reveal information, rather than direct disclosure, must be addressed. As noted previously, taxes are another important friction that effect dividend policy (for example, see Allen et al. 2000).

Finally, Miller and Upton (1976) show that firms are indifferent between leasing and buying capital, except when they face different tax rates. Myers et al. (1976) develop a formula to evaluate the lease versus buy decision, where different tax rates across firms create different discount rates.

They show it is optimal for low tax rate, and hence high discount rate firms, to lease. Alchian and Demsetz (1972) show that leasing involves agency costs due to the separation of ownership and control of capital; a lessee may not have the same incentive as an owner to properly use or maintain the capital. Coase (1972) and Bulow (1986) argue that a durable goods monopolist may lease in order to avoid time inconsistency, and Hendel and Lizzari (1999, 2002) show that it may lease to reduce competition or adverse selection in secondary (used goods) markets. Eisfeldt and Rampini (2007) show that leasing has a repossession advantage relative to buying via secured lending. The trade-off involves the benefit of the enforcement advantage for leased capital, relative to the cost of the ownership, versus a standard control agency problem which arises because ownership and control are separated.

In addition to these specific advances in financial structure, an essential part of Modigliani and Miller's innovation was to put agents on an equal footing. They, and others, then asked what types of friction would cause agents to have different market opportunities, information sets or commitment frictions? This perspective, which was novel at the time, has been used productively to analyse problems in monetary economics, public finance, international economics, and a number of other applications. In summary, the most profound and lasting impacts of the Modigliani–Miller theorem have been this notion of 'even footedness' and the systematic investigation of the theorem's assumptions. The approach has motivated decades of research in economics and finance in a search for what *is* relevant in a host of economic problems (between borrowers and lenders, governments and citizens, and countries). As Miller (1988, p. 100) said: 'Showing what doesn't matter can also show, by implication, what does.'

See Also

- ▶ Arbitrage
- ▶ Finance
- ▶ Miller, Merton (1923–2000)
- ▶ Modigliani, Franco (1918–2003)

Bibliography

Aghion, P., and P. Bolton. 1992. An incomplete contracts approach to financial contracting. *Review of Economic Studies* 59: 473–494.

Alchian, A., and H. Demsetz. 1972. Production, information costs and economic organization. *American Economic Review* 62: 777–795.

Allen, F., A. Bernardo, and I. Welch. 2000. A theory of dividends based on tax clienteles. *Journal of Finance* 55: 2499–2536.

Allen, F., and D. Gale. 1988. Optimal security design. *Review of Financial Studies* 1: 229–263.

Allen, F., and D. Gale. 1991. Arbitrage, short sales and financial innovation. *Econometrica* 59: 1041–1068.

Bhattacharya, S. 1979. Imperfect information, dividend policy, and the 'bird in the hand' fallacy. *Bell Journal of Economics* 10: 259–270.

Bulow, J. 1986. An economic theory of planned obsolescence. *Quarterly Journal of Economics* 101: 729–749.

Coase, R. 1972. Durability and monopoly. *Journal of Law and Economics* 15: 142–149.

Eisfeldt, A., and A. Rampini. 2007. Leasing, ability to repossess and debt capacity. *Review of Financial Studies,* forthcoming.

Grossman, S., and O. Hart. 1988. One share one vote and the market for corporate control. *Journal of Financial Economics* 20: 175–202.

Harris, M., and A. Raviv. 1988. Corporate governance: Voting rights and majority rule. *Journal of Financial Economics* 20: 203–235.

Hendel, I., and A. Lizzari. 1999. Interfering with secondary markets. *RAND Journal of Economics* 30: 1–21.

Hendel, I., and A. Lizzari. 2002. The role of leasing under adverse selection. *Journal of Political Economy* 110: 113–143.

Krasa, S., and A.P. Villamil. 2000. Optimal contracts when enforcement is a decision variable. *Econometrica* 68: 119–134.

Lacker, J., and J. Weinberg. 1989. Optimal contracts under costly state falsification. *Journal of Political Economy* 97: 1345–1363.

Leland, H., and D. Pyle. 1977. Informational asymmetries, financial structure and financial intermediation. *Journal of Finance* 32: 371–387.

Miller, M.H. 1977. Debt and taxes. *Journal of Finance* 32: 261–275.

Miller, M.H. 1988. The Modigliani–Miller proposition after thirty years. *Journal of Economic Perspectives* 2(4): 99–120.

Miller, M.H. 1991. *Financial innovations and market volatility.* Cambridge, MA: Blackwell.

Miller, M.H., and F. Modigliani. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* 34: 411–433.

Miller, M.H., and C. Upton. 1976. Leasing, buying and the cost of capital services. *Journal of Finance* 31: 761–786.

Modigliani, F. 1980. Introduction. In *The collected papers of Franco Modigliani*, ed. A. Abel, Vol. 3. Cambridge, MA: MIT Press.
Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporate finance and the theory of investment. *American Economic Review* 48: 261–297.
Modigliani, F., and M.H. Miller. 1963. Corporate income taxes and the cost of capital: A correction. *American Economic Review* 53: 433–443.
Myers, S., D. Dill, and A. Bautista. 1976. Valuation of financial lease contracts. *Journal of Finance* 31: 799–819.
Myers, S., and N. Majluf. 1984. Corporate financing and investment when firms have information that investors do not have. *Journal of Financial Economics* 11: 187–221.
Stiglitz, J. 1969. A re-examination of the Modigliani–Miller theorem. *American Economic Review* 59: 784–793.
Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 22: 265–293.
Zender, J. 1991. Optimal financial instruments. *Journal of Finance* 46: 1645–1663.

Molinari, Gustave de (1819–1912)

R. F. Hébert

Belgian economist and journalist, the most extreme member of the French Liberal School, Molinari was born at Liège in 1819 and died at Paris almost a century later, in 1912. Son of a physician who served as a field officer in the Belgian army, he spent most of his adult life in France, save for a short period after the coup d'état of 1851, when he returned to Belgium to teach economics at the Royal Brussels Museum of Industry and at the Institut Supérieur du Commerce (Antwerp).

Molinari made contact with the Paris group of economists around 1840. He was present at the first meeting of the Société d'Economie Politique and thereafter continued as a member of its inner circle. His steadfast support of the society and his succession to the editorship of the *Journal des Economistes* on Garnier's death in 1881 provided continuity for the liberal viewpoint, of which Molinari was the most adamant spokesman. Trusting completely in free competition and the play of natural forces to solve every social and economic problem, he was optimistic to the verge of utopianism. His conception of the state was narrower than that of the Physiocrats. For example, he denied government the right of expropriation, education and currency issue. Indefatigable in the defence of liberty, he waged a lifelong battle against all forms of government interference, constantly submitting every social action to the sway of three basic laws: self-interest, competition and value. As against the socialist solution, he advocated extension of the corporate form of business organization to further diffuse ownership of property, and the creation of international labour exchanges (similar to their commodity counterparts) to improve employment information and labour mobility.

Molinari's extremism notwithstanding, he typifies a situation in 19th-century economics that was peculiarly French, one that inhibited the emergence of new ideas. The Liberal School exercised a tight monopoly over the institutions of higher learning (the universities and the Institute), the leading professional organization and its journal, as well as the largest publisher of economics books (Guillaumin & Cie). This orthodoxy brooked no opposition and repelled all theoretic novelty. Its intolerance of mathematical economics, for example, drove Walras into academic exile in Switzerland. It is more than slightly ironic that such unvarnished dogmatism acquired the name 'liberal'. Despite its publicity and its dominance, however, the Liberal School did not eclipse all theoretical economics in France. The best analytical work of the period was carried out at the *grandes écoles,* especially at the Ecole des Ponts et Chaussées, training ground of Isnard, Dupuit, Cheysson and others.

Selected Works

1846. *Etudes économiques.* Paris: Capelle.
1863. *Cours d'économie politique*, 2 vols. Paris: Guillaumin.

1880. *L'évolution économique du dix-neuvième siècle. Théorie du progrès*. Paris: C. Reinwald.
1887. *Les lois naturelles de l'économie politique*. Paris: Guillaumin.
1893. *Précis d'économie politique et de morale*. Paris: Guillaumin.
1908. *Economie de l'histoire. Théorie de l'évolution*. Paris: F. Alcan.

References

Guyot, Y. 1912. Gustave de Molinari. *Economic Journal* 22: 152–156.
Pirou, G. 1925. *Les doctrines économiques en France depuis 1870*. Paris: A. Colin.

Monetarism

Phillip Cagan

Keywords

Brunner, K.; Budget deficits; Cash balance approach; Central banks; Chicago School; Consumer expenditure; Consumption multiplier; Crowding out; Friedman, M.; Gold standard; Inflation; Keynesianism; Liquidity trap; Meltzer, A.; Mints, L.; Monetarism; Monetary approach to the balance of payments; Monetary base; Monetary targeting; Monetary transmission mechanism; Money supply; Natural rate of unemployment; Phillips curve; Quantity theory of money; Rational expectations; Simons, H.; Stabilization; Unemployment; Velocity of circulation

JEL Classifications

E5

Monetarism is the view that the quantity of money has a major influence on economic activity and the price level and that the objectives of monetary policy are best achieved by targeting the rate of growth of the money supply.

Background and Initial Development

Monetarism is most closely associated with the writings of Milton Friedman who advocated control of the money supply as superior to Keynesian fiscal measures for stabilizing aggregate demand. Friedman (1948) had proposed that the government finance budget deficits by issuing new money and use budget surpluses to retire money. The resulting countercyclical variations in the money stock would stabilize the economy, provided that the government set its expenditures and tax rates to balance the budget at full employment. In his *A Program for Monetary Stability* (1960), however, Friedman proposed that constant growth of the money stock, divorced from the government budget, would be simpler and equally effective for stabilizing the economy.

In their emphasis on the importance of money, these proposals followed a tradition of the Chicago School of economics. Preceding Friedman at the University of Chicago, Henry Simons (1936) had advocated control of the money stock to achieve a stable price level, and Lloyd Mints (1950) laid out a specific monetary programme for stabilizing an index of the price level. These writers rejected reliance on the gold standard because it had failed in practice to stabilize the price level or economic activity. Such views were not confined to the University of Chicago. In the 1930s James Angell of Columbia University (1933) advocated constant monetary growth, and in the post-Second World War period Karl Brunner and Allan Meltzer were influential proponents of monetarism. The term 'monetarism' was first used by Brunner (1968). He and Meltzer founded the 'Shadow Open Market Committee' in the 1970s to publicize monetarist views on how the Federal Reserve should conduct monetary policy. Monetarism gradually gained adherents not only in the United States but also in Britain (Laidler 1978) and other Western European countries, and subsequently around the world. The growing prominence of monetarism led to intense controversy among economists over the desirability of a policy of targeting monetary growth.

The roots of monetarism lie in the quantity theory of money which formed the basis of classical monetary economics from at least the 18th century. The quantity theory explains changes in nominal aggregate expenditures – reflecting changes in both the physical volume of output and the price level – in terms of changes in the money stock and in the velocity of circulation of money (the ratio of aggregate expenditures to the money stock). Over the long run changes in velocity are usually smaller than those in the money stock and in part are a result of prior changes in the money stock, so that aggregate expenditures are determined largely by the latter. Moreover, over the long run growth in the physical volume of output is determined mainly by real (that is, non-monetary) factors, so that monetary changes mainly influence the price level. The observed long-run association between money and prices confirms that inflation results from monetary overexpansion and can be prevented by proper control of the money supply. This is the basis for Friedman's oft-repeated statement that inflation is always and everywhere a monetary phenomenon.

The importance of monetary effects on price movements had been supported in empirical studies by classical and neoclassical economists such as Cairnes, Jevons and Cassel. But these studies suffered from limited data, and the widespread misinterpretation of monetary influences in the Great Depression of the 1930s fostered doubts about their importance in business cycles. As Keynesian theory revolutionized thinking in the late 1930s and 1940s, it offered an influential alternative to monetary interpretations of business cycles.

The first solid empirical support for a monetary interpretation of business cycles came in a series of studies of the United States by Clark Warburton (for example, 1946). Subsequently Friedman and Anna J. Schwartz compiled new data at the National Bureau of Economic Research in an extension of Warburton's work. In 1962 they demonstrated that fluctuations in monetary growth preceded peaks and troughs of all US business cycles since the Civil War. Their dates for significant steps to higher or lower rates of monetary growth showed a lead over corresponding business cycle turns on the average by about a half year at peaks and by about a quarter year at troughs, but the lags varied considerably. Other studies have found that monetary changes take one to two years or more to affect the price level.

In *A Monetary History of the United States, 1867–1960* (1963b) Friedman and Schwartz detailed the role of money in business cycles and argued in particular that severe business contractions like that of 1929–33 were directly attributable to unusually large monetary contractions. Their monetary studies were continued in *Monetary Statistics of the United States* (1970) and *Monetary Trends in the United States and the United Kingdom* (1982). A companion National Bureau study *Determinants and Effects of Changes in the Stock of Money* (1965) by Phillip Cagan presented evidence that the reverse effect of economic activity and prices on money did not account for the major part of their observed correlation, which therefore pointed to an important causal role of money.

The monetarist proposition that monetary changes are responsible for business cycles was widely contested, but by the end of the 1960s the view that monetary policy had important effects on aggregate activity was generally accepted. The obvious importance of monetary growth in the inflation of the 1970s restored money to the centre of macroeconomics.

Monetarism Versus Keynesianism

Monetarism and Keynesianism differ sharply in their research strategies and theories of aggregate expenditures. The Keynesian theory focuses on the determinants of the components of aggregate expenditures and assigns a minor role to money holdings. In monetarist theory money demand and supply are paramount in explaining aggregate expenditures.

To contrast the Keynesian and monetarist theories, Friedman and David Meiselman (1963) focused on the basic hypothesis about economic behaviour underlying each theory: for the Keynesian theory the consumption multiplier

posits a stable relationship between consumption and income, and for the monetarist theory the velocity of circulation of money posits a stable demand function for money. Friedman and Meiselman tested the two theories empirically using US data for various periods by relating consumption expenditures in one regression to investment expenditures, assuming a constant consumption multiplier, and in a second regression to the money stock, assuming a constant velocity. They reported that the monetarist regression generally fitted the data much better. These dramatic results were not accepted by Keynesians, who argued that the Keynesian theory was not adequately represented by a one-equation regression and that econometric models of the entire economy, based on Keynesian theory, were superior to small-scale models based solely on monetary changes.

The alleged superiority of Keynesian models was contested by economists at the Federal Reserve Bank of St Louis (see Andersen and Jordan 1968). They tested a 'St Louis equation' in which changes in nominal GNP depended on current and lagged changes in the money stock, current and lagged changes in government expenditures, and a constant term reflecting the trend in monetary velocity. When fitted to historical US data, the equation showed a strong permanent effect of money on GNP and a weak transitory (and in later work, non-existent) effect of the fiscal variables, contradicting the Keynesian claim of the greater importance of fiscal than monetary policies. Although the St Louis equation was widely criticized on econometric issues, it was fairly accurate when first used in the late 1960s to forecast GNP, which influenced academic opinion and helped bring monetarism to the attention of the business world.

Although budget deficits and surpluses change interest rates and thus can affect the demand for money, monetarists believe that fiscal effects on aggregate demand are small because of the low interest elasticity of money demand. Government borrowing crowds out private borrowing and associated spending, and so deficits have little net effect on aggregate demand. The empirical results of the St Louis equation are taken as confirmation of weak transitory effects. The debate over the effectiveness of fiscal policy as a stabilization tool has produced a large literature.

In their analysis of the transmission of monetary changes through the economy, Brunner and Meltzer (1976) compare the effects of government issues of money and bonds. If the government finances increased expenditures in a way that raises the money supply, aggregate expenditures increase and nominal income rises. Moreover, the increased supply of money adds to the public's wealth, and greater wealth increases the demand for goods and services. This too raises nominal income. The rise in nominal income is at first mainly a rise in real income and later a rise in prices. They compare this result with one in which the government finances its increased expenditures by issuing bonds rather than money. Again wealth increases, and this raises aggregate expenditures. As long as the government issues either money or bonds to finance a deficit, nominal income must rise due to the increase in wealth. Brunner and Meltzer therefore agree with Keynesians that in principle a deficit financed by bonds as well as by new money is expansionary. However, they show that the empirical magnitudes of the economy are such that national income rises more from issuing a dollar of money than a dollar of bonds.

Policy Implications of Monetarism

Because monetary effects have variable lags of one to several quarters or more, countercyclical monetary policy actions are difficult to time properly. Friedman as well as Brunner and Meltzer argued that an active monetary policy, in the absence of an impossibly ideal foresight, tends to exacerbate, rather than smooth, economic fluctuations. In their view a stable monetary growth rate would avoid monetary sources of economic disturbances, and could be set to produce an approximately constant price level over the long run. Remaining instabilities in economic activity would be minor and, in any event, were beyond the capabilities of policy to prevent. A commitment by the monetary authorities to stable

monetary growth would also help deflect constant political pressures for short-run monetary stimulus and would remove the uncertainty for investors of the unexpected effects of discretionary monetary policies.

A constant monetary growth policy can be contrasted with central bank practices that impart pro-cyclical variations to the money supply. It is common for central banks to lend freely to banks at times of rising credit demand in order to avoid increases in interest rates. Although such interest-rate targeting helps to stabilize financial markets, the targeting often fails to allow rates to change sufficiently to counter fluctuations in credit demands. By preventing interest rates from rising when credit demands increase, for example, the policy leads to monetary expansion that generates higher expenditures and inflationary pressures. Such mistakes of interest-rate targeting were clearly demonstrated in the 1970s, when for some time increases in nominal interest rates did not match increases in the inflation rate, and the resulting low rates of interest in real terms (that is, adjusted for inflation) overstimulated investment and aggregate demand.

The same accommodation of market demands for bank credit results from the common practice of targeting the volume of borrowing from the central bank. Attempts to keep this volume at some designated level require the central bank to supply reserves through open market operations as an alternative to borrowing by banks when rising market credit demands tighten bank reserve positions, and to withdraw reserves in the opposite situation. The resulting procyclical behaviour of the money supply could be avoided by operations designed to maintain a constant growth rate of money.

Brunner and Meltzer (1964a) developed an analytic framework describing how monetary policy should aim at certain intermediate targets as a way of influencing aggregate expenditures. The intermediate targets are such variables as the money supply or interest rates. (Since the Federal Reserve does not control long-term interest rates or the money stock directly, it operates through instrumental variables, such as bank reserves or the federal funds rate, which it can affect directly.) The question of the appropriate intermediate targets of monetary policy soon became the most widely discussed issue in monetary policy.

In recognition of the deficiencies of interest-rate targeting, some countries turned during the 1970s to a modified monetary targeting in which annual growth ranges were announced and adhered to, though with frequent exceptions to allow for departures deemed appropriate because of disturbances from foreign trade and other sources. Major countries adopting some form of monetary targeting included the Federal Republic of Germany, Japan, and Switzerland, all of which kept inflation rates low and thus advertised by example the anti-inflationary virtues of monetarism. In the United States the Federal Reserve also began to set monetary target ranges during the 1970s but generally did not meet them and continued to target interest rates. In October 1979, when inflation was escalating sharply, the Federal Reserve announced a more stringent targeting procedure for reducing monetary growth. Although the average growth rate was reduced, the large short-run fluctuations in monetary growth were criticized by monetarists. In late 1982 the Federal Reserve relaxed its pursuit of monetary targets.

By the mid-1980s the US and numerous other countries were following a partial form of monetary targeting, in which relatively broad bands of annual growth rates are pursued but still subject to major departures when deemed appropriate. These policies are monetarist only in the sense that one or more monetary aggregates are an important indicator of policy objectives; they fall short of a firm commitment to a steady, let alone a non-inflationary, monetary growth rate.

Monetarist Theory

Monetarist theory of aggregate expenditures is based on a demand function for monetary assets that is claimed to be stable in the sense that successive residual errors are generally offsetting and do not accumulate. Given the present inconvertible-money systems, the stock of money is treated as under the control of the government.

Although a distinction is made in theory between the determinants of household and business holdings of money, money demand is usually formulated for households and applied to the total. In these formulations the demand for money depends on the volume of transactions, the fractions of income and of wealth the public wishes to hold in the form of money balances, and the opportunity costs of holding money rather than other income-producing assets (that is, the difference between yields on money and on alternative assets). The alternative assets are viewed broadly to include not only financial instruments but also such physical assets as durable consumer goods, real property, and business plant and equipment. The public is presumed to respond to changes in the amount of money supplied by undertaking transactions to bring actual holdings of both money and other assets into equilibrium with desired holdings. As a result of substitutions between money and assets, starting with close substitutes, yields change on a broad range of assets, including consumer durables and capital goods, in widening ripples that affect borrowing, investment, consumption, and production throughout the economy.

The end result is reflected in *aggregate* expenditures and the average level of prices. Independently of this monetary influence on aggregate expenditures and the price level, developments specific to particular sectors determine the distribution of expenditures among goods and services and relative prices. Thus monetarist theory rejects the common technique for forecasting aggregate output by adding up the forecasts for individual industries or the common practice of explaining changes in the price level in terms of price changes for particular goods and services.

Monetarists were early critics of the once influential Keynesian theory of a highly elastic demand for money with respect to short-run changes in the interest rate on liquid short-term assets, which in extreme form became a 'liquidity trap'. Empirical studies have found instead that interest rates on savings deposits and on short-term market securities have elasticities smaller even than the $-\frac{1}{2}$ implied by the simple Baumol–Tobin cash balance theory (Baumol 1952; Tobin 1956).

In empirical work a common form of the demand function for money includes one or two interest rates and real GNP as a proxy for real income. A gradual adjustment of actual to desired money balances is allowed for, implying that a full adjustment to a change in the stock is spread over several quarters. The lagged adjustment is subject to an alternative interpretation in which money demand reflects 'permanent' instead of current levels of income and interest rates. This interpretation de-emphasizes the volume of transactions as the major determinant of money demand in favour of the monetarist view of money as a capital asset yielding a stream of particular services and dependent on 'permanent' values of wealth, income, and interest rates (in most studies captured empirically by a lagged adjustment). Treatment of the demand for money as similar to demands for other assets stocks is now standard practice.

The monetarist view of money as a capital asset suggests that the demand for it depends on a variety of characteristics, and not uniquely on its transactions services. The definition of money for policy purposes depends on two considerations: the ability of the monetary authorities to control its quantity, and the empirical stability of a function describing the demand for it. In their study of the United States Friedman and Schwartz used an early version of M2, which included time and savings deposits at commercial banks, but they argued that minor changes in coverage would not greatly affect their findings. Subsequently the quantity of transaction balances M1 has become the most widely used definition of money for most countries, though many central banks claim to pay attention also to broader aggregates in conducting monetary policy.

In view of the wide range of assets into which the public may shift any excess money balances, the transmission of monetary changes through the economy to affect aggregate expenditures and other variables can follow a variety of paths. Monetarists doubt that these effects can be adequately captured by a detailed econometric model which prescribes a fixed transmission path. Instead they prefer models that dispense with detailed transmission paths and focus on a stable

overall relationship between changes in money and in aggregate expenditures.

In both the monetarist model and large-scale econometric models, changes in the money stock are usually treated as exogenous (that is, as determined outside the model). It is clear that money approaches a strict exogeneity only in the long run. The US studies by Friedman and Schwartz and by Cagan established that the money supply not only influences economic activity but also is influenced by it in turn. This creates difficulties in testing empirically for the monetary effects on activity because allowance must be made for the feedback effect of economic activity on the money supply. Econometric models of the money supply can allow for feedback through the banking system (Brunner and Meltzer 1964b). Under modern systems of inconvertible money, however, the feedback is dominated by monetary policies of the central banks, and attempts to model central bank behaviour have been less than satisfactory. Statistical tests of the exogeneity of the money supply using the Granger–Sims methodology have given mixed results. Although the concurrent mutual interaction between money and economic activity remains difficult to disentangle, the longer the lag in monetary effects the less likely that the feedback from activity to money can account for the observed association. In the St Louis equation, for example, while the correlation between changes in GNP and in money concurrently could largely reflect feedback from GNP to money, the correlation between changes in GNP and lagged changes in money are less likely to be dominated by such feedback.

Opposition to Monetary Targeting

While monetarism has refocused attention on money and monetary policy, there is widespread doubt that velocity is sufficiently stable to make targeting of monetary growth desirable. Movements in velocity when monetary growth is held constant produce expansionary and contractionary effects on the economy. In the United States the trend of velocity was fairly stable and predictable from the early 1950s to the mid-1970s, but money demand equations based on that period showed large overpredictions after the mid-1970s (Judd and Scadding 1982). Financial innovations providing new ways of making payments and close substitutes for holding money were changing the appropriate definition of money and the parameters of the demand function. In the United States the gradual removal of ceilings on interest rates banks could pay on deposits played a major role in these developments by increasing competition in banking. In Great Britain the removal of domestic controls over international financial transactions led to unusual movements in money holdings in 1979–80. Germany and Switzerland also found growing international capital inflows at certain times a disruptive influence on their monetary policies.

The ‘monetary theory of the balance of payments’ (Frenkel and Johnson 1976) is an extension of monetarism to open economies where money supply and demand are interrelated among countries through international payments. A debated issue is whether individual countries, even under flexible exchange rates, can pursue largely independent monetary policies. The growing internationalization of capital markets is often cited as an argument against the monetarist presumption that velocity and the domestic money supply under flexible foreign exchange rates are largely independent of foreign influences.

Uncertainties over the proper definition of money and instability in the velocity of money as variously defined led to monetarist proposals to target the monetary liabilities of the central bank, that is, the ‘monetary base’ consisting of currency outstanding and bank reserves. The monetary base has the advantage of not being directly affected by market innovations and so of not needing redefinitions when innovations occur. Monetarists have proposed maintaining a constant growth rate of the base also because it would simplify – indirectly virtually eliminate – the monetary policy function of central banks and governments. Some of the European central banks have found targeting the monetary base preferable to targeting the money supply, though

not without important discretionary departures from the target.

Yet financial market developments can also produce instabilities in the relationship between the monetary base and aggregate expenditures. Economists opposed to monetarism propose instead that stable growth of aggregate expenditures be the target of monetary policy and that it be pursued by making discretionary changes as deemed appropriate in growth of the base. This contrasts sharply with the monetarist opposition to discretion in the conduct of policy.

The Phillips Curve Trade-Off

The inflationary outcome of discretionary monetary policy since the Second World War can be explained in terms of the Phillips curve trade-off between inflation and unemployment. Along the Phillips curve lower and lower unemployment levels are associated with higher and higher inflation rates. Such a relationship, first found in historical British data, was shown to fit US data for the 1950s and 1960s and earlier. The trade-off depends on sticky wages and prices. As aggregate demand increases, the rise in wages and prices trails behind, inducing an expansion of output to absorb part of the increase in demand. US experience initially suggested that any desired position on the Phillips curve could be maintained by the management of aggregate demand. Thus a lower rate of unemployment could be achieved and maintained by tolerating an associated higher rate of inflation. Given this presumed trade-off, policymakers tended to favour lower unemployment at the cost of higher inflation.

In the 1970s, however, the Phillips curve shifted towards higher rates of inflation for given levels of unemployment. Friedman (1968) argued that the economy gravitates toward a 'natural rate of unemployment' which in the long run is largely independent of the inflation rate and cannot be changed by monetary policy. Wages and prices adjust sluggishly to unanticipated changes in aggregate demand but adjust more rapidly to maintained increases in demand and prices that are anticipated. Consequently, the only way to hold unemployment below the natural rate is to keep aggregate demand rising faster than the anticipated rate of inflation. Since the anticipated rate tends to follow the actual rate upward, this leads to faster and faster inflation. This 'acceleration principle' implies that there is no permanent trade-off between inflation and unemployment. The existence of a natural rate of unemployment also implies that price stability does not lead to higher unemployment in the long run.

Monetarist thought puts primary emphasis on the long-run consequences of policy actions and procedures. It rejects attempts to reduce short-run fluctuations in interest rates and economic activity as usually beyond the capabilities of monetary policy and as generally inimical to the otherwise achievable goals of long-run price stability and maximum economic growth. Monetarists believe that economic activity, apart from monetary disturbances, is inherently stable. Much of their disagreement with Keynesians can be traced to this issue.

Rational Expectations

One version of the rational expectations theory goes beyond monetarism by contending that there is little or no Phillips curve trade-off between inflation and unemployment even in the short run, since markets are allegedly able to anticipate any systematic countercyclical policy pursued to stabilize the economy. Only unanticipated departures from such stabilization policies affect output; all anticipated monetary changes are fully absorbed by price changes. Since unsystematic policies would have little countercyclical effectiveness or purpose, the best policy is to minimize uncertainty with a predictable monetary growth.

This theory shares the monetarist view that unpredictable fluctuations in monetary growth are an undesirable source of uncertainty with little benefit. But the two views disagree on the speed of price adjustments to predictable monetary measures and on the associated effects on economic activity. Monetarists do not claim that countercyclical policies have no real effects, but they are sceptical of our ability to use them effectively. It is

the ill-timing of countercyclical policies as a result of variable lags in monetary effects that underlies the monetarist preference for constant monetary growth to avoid uncertainty and inflation bias.

Interest in Private Money Supplies

Monetarism is the fountainhead of a renewed interest in a subject neglected during the Keynesian Revolution: the design of monetary systems that maintain price-level stability. Scepticism that price-level stability can be achieved even by a constant growth rate of money however defined or of the monetary base has led to proposals for a strict gold standard or for a monetary system in which money is supplied by the private sector under competitive pressures to maintain a stable value. While monetarists are sympathetic to proposals to eliminate discretionary monetary policies, they view such alternative systems as impractical and believe that a nondiscretionary government policy of constant monetary growth is the best policy.

Associated Views of the Monetarist School

Monetarism is associated with various related attitudes towards government (see Mayer 1978). Monetarism shares with laissez-faire a belief in the long-run benefits of a competitive economic system and of limited government intervention in the economy. It opposes constraints on the free flow of credit and on movements of interest rates, such as the US ceilings on deposit interest rates (removed by the mid- 1980s except on demand deposits). The disruptive potential of such ceilings became evident in the 1970s when financial innovations, partly undertaken to circumvent the ceilings, produced the transitional shifts in the traditional money-demand functions that created difficulties for the conduct of monetary policy. Government control over the quantity of money is viewed as a justifiable exception to laissez-faire, however, in order to ensure the stability of the value of money.

See Also

- ▶ Friedman, Milton (1912–2006)
- ▶ Keynesianism
- ▶ Monetary Policy, History of
- ▶ New Classical Macroeconomics
- ▶ Quantity Theory of Money
- ▶ Rational Expectations

Bibliography

Andersen, L.C., and J.L. Jordan. 1968. Monetary and fiscal actions: A test of their relative importance in economic stabilization. *Federal Reserve Bank of St Louis Review* 50: 11–24.

Angell, J. 1933. Monetary control and general business stabilization. In *Economic essays in honour of Gustav Cassel*. London: Allen and Unwin.

Baumol, W.J. 1952. The transactions demand for cash: An inventory theoretic approach. *Quarterly Journal of Economics* 66: 545–556.

Brunner, K. 1968. The role of money and monetary policy. *Federal Reserve Bank of St Louis Review* 50: 8–24.

Brunner, K. and A. Meltzer 1964a. The Federal Reserve's attachment to the free reserve concept. U.S. Congress House Committee on Banking and Currency, Subcommittee on Domestic Finance, April.

Brunner, K., and A. Meltzer. 1964b. Some further investigations of demand and supply functions for money. *Journal of Finance* 19: 240–283.

Brunner, K., and A. Meltzer. 1976. An aggregative theory for a closed economy. In *Studies in monetarism*, ed. J. Stein. Amsterdam: North-Holland.

Cagan, P. 1965. *Determinants and effects of changes in the stock of money 1875–1960*. New York: Columbia University Press.

Frenkel, J.A., and H.G. Johnson, eds. 1976. *The monetary approach to the balance of payments*. Toronto: University of Toronto Press.

Friedman, M. 1948. A monetary and fiscal framework for economic stability. *American Economic Review* 38: 256–264.

Friedman, M. 1960. *A program for monetary stability*. New York: Fordham University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Friedman, M., and D. Meiselman. 1963. The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958. In *Commission on money and credit*, *Stabilization policies*. Englewood Cliffs: Prentice-Hall.

Friedman, M., and A.J. Schwartz. 1963a. Money and business cycles. *Review of Economics and Statistics* 45 (1): 32–64. Part II, Supplement.

Friedman, M., and A. Schwartz. 1963b. *A monetary history of the United States 1867–1960*. Princeton: Princeton University Press.
Friedman, M., and A. Schwartz. 1970. *Monetary statistics of the United States estimates, sources, methods*. New York: NBER.
Friedman, M., and A. Schwartz. 1982. *Monetary trends in the United States and the United Kingdom their relation to income, prices and interest rates, 1867–1975*. Chicago: University of Chicago Press.
Judd, J.P., and J.L. Scadding. 1982. The search for a stable money demand function: A survey of the post-1973 literature. *Journal of Economic Literature* 20: 993–1023.
Laidler, D. 1978. Mayer on monetarism: Comments from a British point of view. In *The structure of monetarism*, ed. T. Mayer. New York: Norton.
Mayer, T., ed. 1978. *The structure of monetarism*. New York: Norton.
Mints, L.W. 1950. *Monetary policy for a competitive society.* New York: McGrawHill.
Simons, H. 1936. Rules versus authorities in monetary policy. *Journal of Political Economy* 44: 1–30.
Tobin, J. 1956. The interest elasticity of transactions demand for cash. *Review of Economics and Statistics* 38: 241–247.
Warburton, C. 1946. The misplaced emphasis in contemporary business-fluctuation theory. *Journal of Business* 19: 199–220.

Monetary Aggregation

William A. Barnett

Abstract

Aggregation theory and index-number theory provide the foundations for official governmental data. However, the monetary quantity aggregates and interest rate aggregates supplied by many central banks are not based on index-number or aggregation theory, but rather are the simple unweighted sums of the component quantities and the quantity-weighted or unweighted arithmetic averages of interest rates. The result has been instability of estimated money demand and supply functions, and a series of 'puzzles' in the related applied literature. In contrast, the Divisia monetary aggregates are derived directly from economic index-number theory.

Keywords

Aggregation theory; Barnett critique; Data construction; Equity premium puzzle; European Central Bank; Divisia index; Index number theory; Inflation targeting; Interest rate targeting; Intermediate targets; Monetary aggregation; Monetary policy; Monetary quantity targeting; Monetary targeting; Monetary velocity; Money demand; Money supply

JEL Classifications

C43; E51; E41; G12; C43; C22

Aggregation theory and index-number theory have been used to generate official governmental data since the 1920s. One exception still exists. The monetary quantity aggregates and interest rate aggregates supplied by many central banks are not based on index-number or aggregation theory, but rather are the simple unweighted sums of the component quantities and quantity-weighted or arithmetic averages of interest rates. The predictable consequence has been induced instability of money demand and supply functions, and a series of 'puzzles' in the resulting applied literature. In contrast, the Divisia monetary aggregates, originated by Barnett (1980), are derived directly from economic index-number theory. Financial aggregation and index number theory was first rigorously connected with the literature on microeconomic aggregation and index number theory by Barnett (1980, 1987). A collection of many of his contributions to that field is available in Barnett and Serletis (2000).

Data construction and measurement procedures imply the theory that can rationalize the procedure. The assumptions implicit in the data construction procedures must be consistent with the assumptions made in producing the models within which the data are to be used. Unless the theory is internally consistent, the data and its applications are incoherent. Without that coherence between aggregator function structure and the econometric models within which aggregates are embedded, stable structure can appear to be unstable. This phenomenon has been

called the 'Barnett critique' by Chrystal and MacDonald (1994).

Aggregation Theory Versus Index Number Theory

The exact aggregates of microeconomic aggregation theory depend on unknown aggregator functions, which typically are utility, production, cost, or distance functions. Such functions must first be econometrically estimated. Hence the resulting exact quantity and price indexes become estimator- and specification-dependent. This dependency is troublesome to governmental agencies, which therefore view aggregation theory as a research tool rather than a data construction procedure.

Statistical index-number theory, on the other hand, provides indexes which are computable directly from quantity and price data, without estimation of unknown parameters. Such index numbers depend jointly on prices and quantities, but not on unknown parameters. In a sense, index number theory trades joint dependency on prices and quantities for dependence on unknown parameters. Examples of such statistical index numbers are the Laspeyres, Paasche, Divisia, Fisher ideal, and Törnqvist indexes.

The loose link between index number theory and aggregation theory was tightened, when Diewert (1976) defined the class of second-order 'superlative' index numbers. Statistical index number theory became part of microeconomic theory, as economic aggregation theory had been for decades, with statistical index numbers judged by their nonparametric tracking ability to the aggregator functions of aggregation theory.

For decades, the link between statistical index number theory and microeconomic aggregation theory was weaker for aggregating over monetary quantities than for aggregating over other goods and asset quantities. Once monetary assets began yielding interest, monetary assets became imperfect substitutes for each other, and the 'price' of monetary-asset services was no longer clearly defined. That problem was solved by Barnett (1978, 1980), who derived the formula for the user cost of demanded monetary services. Subsequently Barnett (1987) derived the formula for the user cost of supplied monetary services. A regulatory wedge can exist between the demand and supply-side user costs if non-payment of interest on required reserves imposes an implicit tax on banks.

Barnett's results on the user cost of the services of monetary assets set the stage for introducing index number theory into monetary economics.

The Economic Decision

Consider a decision problem over monetary assets that illustrates the capability of monetary aggregation theory. The decision problem will be defined so that the relevant literature on economic aggregation over goods is immediately applicable. Initially we shall assume perfect certainty.

Let $\mathbf{m}_t' = (m_{1t}, m_{2t}, \ldots m_{nt})$ be the vector of real balances of monetary assets during period t, let r_t be the vector of nominal holding-period yields for monetary assets during period t, and let R_t be the one-period holding yield on the benchmark asset during period t. The benchmark asset is defined to be a pure investment that provides no services other than its yield, R_t, so that the asset is held solely to accumulate wealth. Thus, R_t is the maximum holding period yield in the economy in period t.

Let y_t be the real value of total budgeted expenditure on monetary services during period t. Under simplifying assumptions for data within one country, the conversion between nominal and real expenditure on the monetary services of one or more assets is accomplished using the true cost of living index on consumer goods. But for multi-country data or data aggregated across heterogeneous regions, the correct deflator can be found in Barnett (2003, 2007). The optimal portfolio allocation decision is:

$$\begin{array}{c} \text{maximize } u(\mathbf{m}_t) \\ \text{subject to } \pi_t' \ \mathbf{m}_t = \mathrm{y_t}, \end{array} \tag{1}$$

where $\pi_t' = (\pi_{1t}, \ldots, \pi_{nt})$ is the vector of monetary-asset real user costs, with

$$\pi_{it} = \frac{R_t - r_{it}}{1 + R_t}. \tag{2}$$

This function u is the decision maker's utility function, assumed to be monotonically increasing and strictly concave. The user cost formula (2), derived by Barnett (1978, 1980), measures the forgone interest or opportunity cost of holding monetary asset i, when the higher yielding benchmark asset could have been held.

To be an admissible quantity aggregator function, the function u must be weakly separable within the consumer's complete utility function over all goods and services. Producing a reliable test for weak separability is the subject of much intensive research by an international group of econometricians (see, for example, Jones et al. 2005; Fleissig and Whitney 2003; De Peretti 2005). Two approaches exist. One approach uses stochastic extensions of nonparametric revealed preference tests, while the other uses parametric econometric models.

Let $\mathbf{m}_t^*$ be derived by solving decision (1). Under the assumption of linearly homogeneous utility, the exact monetary aggregate of economic theory is the utility level associated with holding the portfolio, and hence is the optimized value of the decision's objective function:

$$M_t = u\left(\mathbf{m}_t^*\right). \tag{3}$$

The Divisia Index

Although Eq. 3 is exactly correct, it depends upon the unknown function, u. Nevertheless, statistical index-number theory enables us to track M_t exactly without estimating the unknown function, u. In continuous time, the exact monetary aggregate, $M_t = u\left(\mathbf{m}_t^*\right)$, can be tracked exactly by the Divisia index, which solves the differential equation

$$\frac{d\log M_t}{dt} = \sum_i s_{it} \frac{d\log m_{it}^*}{dt} \tag{4}$$

for M_t, where

$$s_{it} = \frac{\pi_{it} m_{it}^*}{y_t}$$

is the i'th asset's share in expenditure on the total portfolio's service flow. In Eq. 4, it is understood that the result is in continuous time, so the time subscripts are a shorthand for functions of time. We use t to be the time period in discrete time, but the instant of time in continuous time. The dual user cost price aggregate $\Pi_t = \Pi\left(\pi_t\right)$, can be tracked exactly by the Divisia price index, which solves the differential equation

$$\frac{d\ \log \Pi_t}{dt} = \sum_i s_{it} \frac{d\ \log \pi_{it}}{dt}. \tag{5}$$

The user cost dual satisfies Fisher's factor reversal in continuous time:

$$\Pi_t M_t = \pi_t' \mathbf{m}_t. \tag{6}$$

As a formula for aggregating over quantities of perishable consumer goods, that index was first proposed by François Divisia (1925) with market prices of those goods inserted in place of the user costs in Eq. 4. In continuous time, the Divisia index, under conventional neoclassical assumptions, is exact. In discrete time, the Törnqvist approximation is:

$$\log M_t - \log M_{t-1} = \sum_i \bar{s}_{it} \left(\log m_{it}^* - \log m_{i,t-1}^*\right), \tag{7}$$

where

$$\bar{s}_{it} = \frac{1}{2}\left(s_{it} + s_{i,t-1}\right).$$

In discrete time, we often call Eq. 7 simply the Divisia quantity index. After the quantity index is computed from (7), the user cost aggregate most commonly is computed directly from Eq. 6.

Diewert (1976) defines a 'superlative index number' to be one that is exactly correct for a quadratic approximation to the aggregator function. The discretization (7) to the Divisia index is in the superlative class, since it is exact for the quadratic translog specification to an aggregator function. With weekly or monthly monetary data, Barnett (1980) has shown that the Divisia index growth rates, (7), are accurate to within three decimal places. In addition, the difference

between the Fisher ideal index and the discrete Divisia index growth rates are third order and comparably small. That third-order differential error typically is smaller than the round-off error in the component data.

Prior Applications

Divisia monetary aggregates were first constructed for the United States by Barnett (1980), when he was on the staff of the Special Studies Section of the Board of Governors of the Federal Reserve System, and are now maintained by the Federal Reserve Bank of Saint Louis in its data base, called FRED (see Anderson et al. 1997, who produced the Divisia data for FRED). A Divisia monetary-aggregates data base also has been produced for the United Kingdom by the Bank of England. An overview of Divisia data maintained by many central banks throughout the world can be found in Belongia and Binner (2000, 2005) and in Barnett et al. (1992), along with a survey of empirical results with that data. The most extensive collection of relevant applied and theoretical research in that area is in Barnett and Serletis (2000) and Barnett and Binner (2004).

The State of the Art

The European Central Bank is implementing a multilateral extension of the Divisia monetary aggregates for monetary quantity and interest rate aggregation within the euro area. This aggregation is multilateral in the recursive sense that it permits aggregation of monetary service flows first within countries, then over countries. The resulting aggregation will be in a strictly nested, internally consistent manner. The multilateral extension of the theory was produced by Barnett (2003, 2007). This extension was produced under three increasingly strong sets of assumptions: (a) with the weakest being produced from heterogeneous agents theory, (b) followed by the somewhat stronger assumption of existence of a multilateral representative agent, and (c) finally with the strongest being the assumption of the existence of a unilateral representative agent. The intent is to move from the weakest towards the strongest assumptions, as progress is made within the European Monetary Union towards its harmonization and economic convergence goals. Since Barnett's three assumption structures are nested, construction of the data under the most general heterogeneous countries approach would continue to be valid, as the stronger assumptions become more reasonable and are attained within the euro area.

Extension of index number theory to the case of risk was introduced by Barnett et al. (2000), who derived the extended theory from Euler equations rather than from the perfect-certainty first-order conditions used in the earlier index number-theory literature. Since that extension is based upon the consumption capital-asset-pricing model (CCAPM), the extension is subject to the 'equity premium puzzle' of smaller than necessary adjustment for risk. We believe that the under-correction produced by CCAPM results from its assumption of intertemporal blockwise strong separability of goods and services within preferences. Barnett and Wu (2005) have extended Barnett, Liu, and Jensen's result to the case of risk aversion with intertemporally non-separable tastes.

The extension to risk is likely to be especially important to countries whose residents hold significant deposits in foreign denominated assets, since exchange-rate risk can cause rates of return on monetary assets to be subject to non-negligible risk. With the recent trend towards financial integration in many parts of the world, exchange-rate risk is likely to grow in importance in monetary aggregation. In many countries, the largest holder of foreign-denominated deposits is the central bank itself. Within the United States, the extension to risk is highly relevant to the so called 'missing M2' episode of the early 1990s, when substitutability among small time deposits, stock funds, and bond funds produced 'puzzles'.

User cost aggregates are duals to monetary quantity aggregates. Either implies the other uniquely. In addition, user-cost aggregates imply the corresponding interest-rate aggregates uniquely. The interest-rate aggregate r_t implied by the user-cost aggregate Π_t is the solution for r_t to the equation:

$$\frac{R_t - r_t}{1 + R_t} = \Pi_t.$$

Accordingly, any monetary policy that operates through the opportunity cost of money (that is, interest rates) has a dual policy operating through the monetary quantity aggregate, and vice versa. Aggregation theory implies no preference for either of the two dual policy procedures or for any other approach to policy, so long as the policy does not violate principles of aggregation theory.

Conclusion

Aggregation theory is about measurement, and has little, if anything, to say about the choice of policy instrument, such as the funds rate or the base. But accurate measurement, through proper application of aggregation theory, has much to say about the transmission of policy, modelling of structure, and the measurement of intermediate targets (if any) and final targets.

Policies that violate aggregation theoretic principles include the following oversimplified approaches: (a) inflation targeting that targets one arbitrary consumer-good price as a final target, while ignoring all other consumer goods prices, rather than targeting the true cost-of-living index over all consumer goods prices; (b) interest rate targeting that analogously targets one arbitrary interest rate as an intermediate target while ignoring all other interest rates, rather than targeting the aggregation-theoretic interest-rate or user-cost aggregate over a weakly separable collection of monetary assets; (c) monetary quantity targeting that targets a simple-sum monetary aggregate as an intermediate target rather than the aggregator function over a weakly separable collection of monetary assets; and (d) policy simulations using money-demand or money-supply functions containing simple-sum monetary aggregates or quantity-weighted interest-rate aggregates. The measurement defects in the above four cases are unrelated to the choice of the funds rate or monetary base as an instrument of policy. Unlike intermediate targets, final targets, and variables in models, the chosen instruments of policy tend to be highly controllable, disaggregated variables, presenting few serious measurement problems.

The objective of the Divisia monetary aggregates is measurement of the economy's monetary service flow and its dual opportunity cost (user cost) and implied interest rate aggregate, not advocacy of any particular policy use of the correctly measured variables. But all uses of data are adversely affected by improper measurement, and a long series of 'puzzles' in monetary economics have been shown to have been produced by improper measurement (see, for example, Barnett and Serletis 2000, ch. 24).

See Also

- European Central Bank
- Federal Reserve System
- Foreign Direct Investment
- Inflation Targeting
- Measurement
- Monetary Economics, History of
- Monetary and Fiscal Policy Overview
- Statistics and Economics

M

Bibliography

Anderson, R., B. Jones, and T. Nesmith. 1997. Building new monetary services indexes: Concepts, data and methods. *Federal Reserve Bank of St Louis Review* 79: 53–82.

Barnett, W. 1978. The user cost of money. *Economics Letters* 1, 145–49. Repinted in Barnett and Serletis (2000, ch. 1).

Barnett, W. 1980. Economic monetary aggregates: An application of aggregation and index number theory. *Journal of Econometrics* 14, 11–48. Reprinted in Barnett and Serletis (2000, ch. 2).

Barnett, W. 1987. The microeconomic theory of monetary aggregation. In *New approaches in monetary economics*, ed. W. Barnett and K. Singleton. Cambridge: Cambridge University Press. Reprinted in Barnett and Serletis (2000, ch. 3).

Barnett, W. 2003. Aggregation-theoretic monetary aggregation over the euro area, when countries are heterogeneous. Working paper no. 260. Frankfurt: European Central Bank.

Barnett, W. 2007. Multilateral aggregation-theoretic monetary aggregation over heterogeneous countries. *Journal of Econometrics* 136: 457–82.

Barnett, W., and J. Binner. 2004. *Functional structure and approximation in econometrics*. Amsterdam: North-Holland.
Barnett, W., and A. Serletis (eds.). 2000. *The theory of monetary aggregation*. Amsterdam: North-Holland.
Barnett, W., and S. Wu. 2005. On user costs of risky monetary assets. *Annals of Finance* 1: 35–50.
Barnett, W., D. Fisher, and A. Serletis. 1992. Consumer theory and the demand for money. *Journal of Economic Literature* 30, 2086–119. Reprinted in Barnett and Serletis (2000, ch. 18).
Barnett, W., Y. Liu, and M. Jensen. 2000. CAPM risk adjustment for exact aggregation over financial assets. *Macroeconomic Dynamics* 1: 485–512.
Belongia, M., and J. Binner. 2000. *Divisia monetary aggregates: Theory and practice*. Basingstoke: Palgrave.
Belongia, M., and J. Binner. 2005. *Money, measurement, and computation*. Basingstoke: Palgrave.
Chrystal, A., and R. MacDonald. 1994. Empirical evidence on the recent behaviour and usefulness of simple-sum and weighted measures of the money stock. *Federal Reserve Bank of St Louis Review* 76: 73–109.
De Peretti, P. 2005. Testing the significance of the departures from utility maximization. *Macroeconomic Dynamics* 9(3): 373–97.
Diewert, W. 1976. Exact and superlative index numbers. *Journal of Econometrics* 4: 115–45.
Divisia, F. 1925. L'Indice monétaire et la théorie de la monnaie. *Revue d'Economie Politique* 39: 980–1008.
Fleissig, A., and G. Whitney. 2003. A new PC-based test for Varian's weak separability conditions. *Journal of Business and Economic Statistics* 21: 133–44.
Jones, B., D. Dutkowsky, and T. Elger. 2005. Sweep programs and optimal monetary aggregation. *Journal of Banking and Finance* 29: 483–508.

Monetary and Fiscal Policy Overview

Narayana R. Kocherlakota

Abstract

This article provides an overview of economic thinking about monetary and fiscal policy. I discuss the methodology of answering policy questions in macroeconomics. I then explain what is known from using this methodology for positive and normative analyses of monetary and fiscal policy. Finally, I describe the challenges associated with endogenizing monetary and fiscal policy.

Keywords

Adjustment costs; Commitment; Computational experimentation; Deflation; Endogenous versus exogenous variables; Game theory; Government budget constraint; Implementability; Incentive compatibility; Inflationary expectations; Lump-sum taxes; Markov-perfect equilibria; Methodology of economics; Mirrlees approach to optimal taxation; Monetary versus fiscal policy; Multiple equilibria; Optimal taxation; Ramsey approach to optimal taxation; Rational expectations; Recursive equilibria; Ricardian equivalence; Social insurance; Sticky prices; Sunspot equilibrium; Taylor rule; Wedges

JEL Classifications

D4; D10

In this article I provide an overview of economic thinking about monetary and fiscal policy. There are three terms that need to be defined in this sentence: policy, monetary, and fiscal. I begin by defining each in turn.

A government's *policy* is akin to a strategy in game theory. It specifies a function at each date that maps the government's information at that date into the government's actions. This information typically takes two forms. First, it includes *endogenous* variables such as past prices, past quantities or past actions of the government. For example, under the famous Taylor rule, a government's choice of current short-term interest rates is based on past observations of the consumer price index and gross domestic product. Second, the government's information includes *exogenous* variables, like the realizations of shocks to productivity or to money demand.

These sources of information may be public or they may be known only to the government. Thus, in the United States the Federal Reserve collects information about the state of the economy that it uses for making decisions but is kept confidential from households in the economy. Note, too, that the government's actions themselves may be private information to the government; for example,

until recently, the Federal Open Market Committee publicly announced its decisions only with a lag.

In the popular press, the term 'policy' is commonly used in a different way, to refer only to the *current* choice of the government. However, as long as *some* economic actors (firms, households or the government itself) are forward-looking, such a specification of policy is intrinsically incomplete. Forward-looking decision-makers need to know not just the government's choice of policy today but also how the government will respond to new information in the future. (This is true even if these forward-looking actors have expectations that are far from rational.) Thus, if the government raises taxes today, my response to that increase depends crucially on whether I believe it will persist for a long time. To make that judgement, I need to know not just the government's choices today but also how its choices in the future depend on new information that the government receives.

Whether a policy is monetary or fiscal or neither depends on the nature of the actions specified by that policy. A policy is said to be *monetary* if the relevant actions are those generally undertaken by a central bank. These may include the size of monetary injections, reserve requirements, the discount rate, or the scale of interventions in bond or foreign exchange markets. A policy is said to be *fiscal* if the relevant actions are tax rates and/or expenditures on various commodities. Of course, many government policies (should Iran be invaded or not?) are neither fiscal nor monetary.

In the body of this article, I discuss several lessons from the study of monetary and fiscal policy. Before doing so, though, it is useful to understand the methodology that was used to learn those lessons (see Lucas 1980, and Prescott 2005, for a fuller discussion of this methodology). Any analysis of policy starts with the following question: on the assumption that no other exogenous variables change, how does the economy respond to a change in policy? This kind of question is really asking about the outcome of a controlled *experiment.* It would be best answered by constructing giant national or super-national laboratories in order to conduct these experiments. But it is clearly impossible to perform controlled experiments of this kind. How then do macroeconomists proceed?

The approach taken by macroeconomists is closely related to the methods used by other non-experimental sciences. Consider for example the issue of global warming. There have been no prior episodes in world history in which man has been able to generate such a large amount of CO_2 in such a short period of time. Hence, there is no way to use prior data to understand the impact of this build-up on climatic variables like temperature. Instead, climatologists rely on computer simulations of abstract models to understand the impact of greenhouse gases on the world's climate.

Similarly, macroeconomists build abstract computational models to answer questions about the impact of monetary and fiscal policy. It is well-understood from many years of computational experimentation that useful models must have certain elements to provide reliable answers to policy questions. The models need to be both dynamic and stochastic in nature. The models need to be explicit about aggregate resource constraints: the amount of goods consumed by governments and households cannot exceed the amount of goods produced. The models should feature households with well-defined objectives and budget constraints. The households and firms in the models should be forward-looking (although they may or may not be fully informed about the state of the economy).

To provide a quantitative answer about the impact of a particular policy, macroeconomists need to be specific about many other elements of the computational model (preferences of households, shocks hitting the economy, and so on). Again, it is useful to refer to the natural sciences as a way to understand how macroeconomists proceed. Consider a biologist that wants to understand the impact of a new drug on human beings. At least initially, she experiments on animals. For some kinds of drugs, she may use mice. For others, she may use more expensive animals like monkeys or dogs. Her decision about which proxy to use is a complex one, grounded in theory, collective prior experience about other drugs and these animals, and individual judgement.

In the same fashion, macroeconomists do not use the same model for all policy questions. Instead, they choose the model based on the question at hand. Thus, for questions concerning the short-run impact of monetary policies, they may include adjustment costs in physical capital and/or prices. For other questions concerning the long-run impact of monetary policy, they may neglect these elements. Like the biologist, their decisions are based on theory, collective prior experience and judgement.

One aspect of this decision-making that receives particular attention in macroeconomics is how to quantify the various elements of the model. How risk-averse are the households in the model economy? What is the elasticity of substitution between capital and labour in the model economy? Fortunately, for many of these parameter choices, there is a profession-wide consensus, informed by many years of experience and discussion. For other parameters, new choices have to be made. Generally, macroeconomists use a mix of information from both microeconomic and macroeconomic sources to make these choices. There may well be a range of plausible choices for a given parameter, and then the answer to the policy question under consideration is really a set, not a single point.

In the remainder of this article, I discuss some of the conclusions about monetary and fiscal policy that macroeconomists have reached from using this methodology. I focus on results that are highly robust, in the sense that they occur across a wide class of models. I begin by looking at lessons from the *positive* approach to policy, which studies the response of the private sector to different specifications of policy. I then look at lessons from the *normative* approach, which looks at properties of *ex ante* optimal policies. Finally, I discuss some difficulties associated with modelling policy choices as being an endogenous response to economic conditions.

The Positive Approach to Policy

There is a large amount of macroeconomic research that treats monetary and fiscal policy as wholly exogenous to the economy. It asks questions of the sort: how does some aspect of private sector economic behaviour respond to a given specification of monetary and fiscal policy? Macroeconomists have described the outcomes to many specific experiments of this kind. There is no useful way to summarize this knowledge. However, there are several general lessons that one can draw from this research. In what follows, I discuss three of these.

Lesson 1. Fiscal Versus Monetary Policy

I have drawn a distinction between fiscal and monetary policy. However, this distinction is more than a little artificial for two reasons. First, in macroeconomic models households face budget constraints and aggregate resource constraints are satisfied. Together, these imply that the government itself must satisfy a budget constraint in equilibrium: the present value of the government's revenues must equal the present value of its expenditures. (There are overlapping-generations model economies in which this restriction need not be satisfied. However, these models are typically not thought to be empirically relevant; Abel et al. 1989.) This constraint implies a sharp linkage between fiscal and monetary policy. Changes in monetary policies affect the government's revenue from money creation. Hence, the two types of policies are inextricably linked, because they cannot be changed separately. (This fundamental linkage between fiscal and monetary policy was made especially clear by Sargent and Wallace 1981.)

The second reason is that, in terms of its impact on the economy, monetary policy is merely fiscal policy by another name. People and firms who hold money are forgoing the interest that they could receive by holding bonds instead. They hold that money because it helps them buy goods and services that are difficult to purchase using bonds. Higher interest rates makes money more costly to hold, and makes those goods and services more costly to buy. The interest rate acts like a sales tax on those goods and services.

Monetary policy has still other distorting effects on the economy when some prices are more flexible than others. For example, suppose nominal wages do not respond rapidly to changes in inflation, but gas prices do. Then, the relative

price of labour and gasoline may vary in response to variations in monetary instruments. Again, though, a particular kind of fiscal policy – variations in the gasoline tax – can affect the economy in exactly the same way (This equivalence between fiscal and monetary policy is stressed by Correia et al. 2004).

Lesson 2. Ricardian Equivalence

I pointed out above that the present value of government expenditures must equal the present value of government revenues. This simple fact has surprising consequences. Consider two policies with the same government purchases. Suppose one policy generates lower tax revenue in the next ten years than the other policy. Obviously, under the first policy, the government must borrow more. This extra demand in loans puts upward pressure on interest rates.

However, the government's intertemporal budget constraint also implies that the first policy must necessarily generate *higher* tax revenue in the future. Forward- looking households anticipate this increase in their future tax burden. They respond by saving more to meet this tax burden. In a classic paper, Barro (1974) shows that, if households are sufficiently forward-looking, and markets are frictionless, then the households' extra demand for savings under the first policy is exactly equal to the government's extra demand for loans. Hence, even though the government is borrowing more, there is no extra pressure on interest rates; they should be the same under the two policies. This result is generally termed *Ricardian equivalence* (because of some antecedents in the work of David Ricardo).

The exact Ricardian equivalence result is not robust to adding plausible frictions like borrowing constraints on households. Nonetheless, there is a qualitative lesson that holds much more generally and is often forgotten in policy discussions: economics does not predict a stable relationship between current government debt or deficits and interest rates.

Lesson 3. Expectations Matter

I have emphasized above that households' expectations about future government actions matter for current outcomes. However, in many macroeconomic models a given household's behaviour depends also on its expectations of other households' current and future actions. This feedback generates the possibility of multiple equilibrium outcomes for a given government policy.

Here's a simple example of this phenomenon. Suppose both government investment and household labour are necessary inputs into production – that is, either zero government investment or zero labour input leads to zero output. Suppose as well that the government collects resources to fund its investment by taxing output.

In such a world, regardless of the government's policy, there is always an equilibrium in which households do not work at all. In this equilibrium, because other households are not working, a given household realizes that the government cannot fund any investment. Hence, it is individually optimal for that household not to provide any labour input.

This kind of multiplicity leads to the possibility of what are called *sunspot* fluctuations in macroeconomic variables. The idea here is that households use some arbitrary random variable to coordinate their behavior. Thus, if they all see rain in Peoria, they decide not to work. If they see sun in Peoria, they decide to work. Whether it is sunny or not in Peoria, of course, is irrelevant for economic fundamentals – but in this economy, this variable can still affect equilibrium outcomes. (For early expositions of the concept of sunspot equilibria, see Azariadis 1981, and Cass and Shell 1983.)

Note that this example is only an illustration of a much more general phenomenon. It is especially prevalent in monetary economies. In these settings, a household's decision about how many real balances to hold today depends crucially on the household's expectations about future inflation rates. Obstfeld and Rogoff (1983) demonstrate how this intertemporal feedback can generate a continuum of welfare-indexed possible inflation paths as equilibria, even if the money supply is fixed. Sargent and Wallace (1975) demonstrate how this intertemporal feedback can generate a continuum of welfare-indexed possible inflation paths as equilibria even if interest rates

are fixed. (Pareto-ranked equilibria do not occur in all economies. In many economies – especially non-monetary ones – it may be possible to prove that any equilibrium allocation solves a maximization problem in which the objective is a weighted average of households' utilities. In such settings, equilibrium allocations are necessarily Pareto non-comparable. Without such a proof in hand, though, one has to be aware that there is the potential for sunspot fluctuations between Pareto- ranked outcomes. Many macroeconomists restrict attention to so-called recursive equilibria or Markov-perfect equilibria. Under these notions of equilibrium, outcomes have the property that they depend on the past only through a small number of state variables. This restriction is undoubtedly useful for simplifying computational or econometric work. However, the restriction may inadvertently rule out important sources of potential multiplicity. See, for example, Woodford's 1994, analysis of Lucas and Stokey's 1987, model economy.)

The Normative Approach to Policy

I now turn to the second approach to studying macroeconomic policy. This approach posits a government that chooses a policy at the beginning of time; its objective is to maximize some weighted average of household utilities. Crucially, the government is able to commit to never change the policy. This kind of commitment power is clearly artificial; the goal of the second approach is to tell us what kinds of policies maximize *ex ante* social welfare, not what policies are actually adopted by governments. By construction, there is no requirement that the optimal policies be realistic: normative analyses tell us what the government should do, not what they actually do. Thus, economists use normative analyses to argue strongly in favour of free trade, which is a policy that has never been followed by any country at any time.

Everything in this approach hinges on what is assumed about the set of instruments available to the government. It is well-known that *lump-sum* taxes are a highly desirable taxation instrument. A lump-sum tax is a tax on a household or firm which is independent of their actions. Such a tax is desirable because it does not distort the choices of the household or the firm.

But lump-sum taxes are typically not used by governments. Once one notices this fact, there are at least two ways to proceed in thinking about optimal taxes. One can assume that the governments can only use a limited set of tax instruments that does not include lump-sum taxes. This approach is generally called the *Ramsey* approach. Alternatively, one can build model economies in which governments have access to all possible tax instruments, but *choose*, because of a particular private information friction, not to use lump-sum taxes. This approach is generally called the Mirrlees approach.

The Ramsey Approach and Its Lessons

Suppose the government can impose a linear tax on capital income, a linear tax on labour income, and can print money. It must optimally choose these instruments so as to finance an optimally chosen process for government purchases. What are the properties of the optimal taxes? An enormous amount of work has been done on this question; see Chari and Kehoe (1999) for a survey. I first briefly describe the mathematical approach, and then turn to the properties of the optimal taxes.

One way to proceed here would be to solve for the households' and firms' response to all possible tax policies. Then, given this response function, we could solve the government's optimization problem. This problem turns out to be difficult in most circumstances.

Fortunately, there is a way to substitute out the tax schedules; we can instead think of the government directly choosing quantities subject to two types of restrictions. The first is the usual physical feasibility constraints. The other is a set of constraints called *implementability* constraints. These look like household budget constraints, except that we substitute the household marginal rates of substitution in for all prices; the constraints then contain only physical quantities. Somewhat remarkably, these simple implementability constraints turn out to capture exactly the seemingly complicated restriction that the government can use only linear taxes.

Of course, because it is couched only in terms of quantities, the solution to this problem does not contain direct information about optimal taxes. Once one solves the optimization problem, one sees that there are differences (commonly termed *wedges*) between marginal rates of substitution and marginal rates of transformation in the solution. The optimal taxes in equilibrium are equal to these wedges from the solution of the optimization problem. Note these wedges exist only because of the implementability constraints; without them, all wedges would be zero, and it would be optimal to set all taxes to zero.

What then are the properties of optimal taxes when we apply this kind of analysis? In general, the quantitative properties of the optimal taxes depend on many precise details of the specification of the environment. However, there are (at least) two remarkably robust properties of the optimal taxes. The first is that if the government can accumulate assets, the long-run optimal capital income tax rate is zero. (This result was originally derived by Chamley 1986, and Judd 1985.) Intuitively, suppose the long-run capital income tax is positive. This tax rate affects the rate of return in every period, and its impact cumulates as the horizon of the investment grows. Hence, the tax rate on accumulating capital between period t and period $t + s$ gets arbitrarily large as t, s get large. This arbitrarily large tax rate creates too much social waste, given that it is raising only a finite amount of revenue. The second property of optimal taxes is that, under very general conditions, the optimal nominal interest rate is zero (in all periods, not just in the long run) (see Chari et al. 1996; Correia and Teles 1999; Correia et al. 2004).

Here, the basic intuition is that any positive nominal interest rate is a tax on money holdings (as discussed above). But money is not a final good; it is only an intermediate input into production and consumption. A tax on intermediate inputs creates two distortions: people are deterred both from using the intermediate input and from consuming any final goods that use the intermediate input. It is generally optimal to eliminate this double distortion by simply taxing final goods and not taxing any intermediate inputs, including money.

Even though the nominal interest rate is zero, the real interest rate can still be positive as long as the price index is falling over time. If prices are fully flexible, then this consistent deflation has no real effects. However, if prices are sticky, this steady deflation may create inefficiencies in a world with sticky prices. In particular, if some prices are adjusted downward more frequently than others, then any consistent deflation creates distortions in relative prices.

Correia et al. (2004)) demonstrate that this kind of systematic distortion can be fixed by using *sales* taxes. Their key observation is that the nominal interest rate can be zero and the real interest rate can be positive as long as the *after-tax* price level is falling over time. Hence, if the government sets the sales tax to fall at the correct rate, firms will find it optimal to never change their prices even though the nominal interest rate is zero.

The Mirrlees Approach and Its Lessons

The Ramsey approach simply assumes that governments cannot use lump-sum taxes. But why do governments not use lump-sum taxes? One problem is that, if the government imposes a tax of, say, $10,000 per head, then some people will have the ability to generate this income and others will not. This is not a difficulty if the government can tell who is in which group – it can just exempt those who cannot pay.

Unfortunately, people can *pretend* to be unable to generate this level of income by pretending to have back pain, mental illness or other sources of disability. The government cannot figure out whom to exempt from the head tax.

This observation suggests that governments are deterred from using lump-sum taxes because people are privately informed about their abilities or skills. The Mirrlees approach starts with this informational restriction. The government is allowed to use any form of taxes that it wishes (linear, nonlinear, and so on) on any private sector choice. Because it is not restricted to linear taxes, the implementability constraint discussed above vanishes. Instead, the government faces an *incentive-compatibility constraint* that reflects the ability of people to pretend to be less able than they truly are.

Given this difference in constraints, one can proceed much as in the Ramsey approach. The first step is to set up a maximization problem in which the government maximizes *ex ante* welfare subject to feasibility constraints and incentive-compatibility constraints. This type of maximization problem is roughly equivalent to the kind of dynamic contracting problems originally considered by Green (1987). One considerable complication is that abilities may change over time due to health shocks. Dynamic contracting models with persistent shocks are highly challenging to solve even with a computer (see Fernandes and Phelan 2000).

The next step is to design a tax system such that the optimal allocations emerge as equilibrium outcomes. These tax systems are complicated objects when abilities evolve over time. Nonetheless, we can draw remarkably strong conclusions about the structure of optimal capital income taxes. If preferences are additively separable between consumption and leisure, then one can show that there exists an optimal tax system which is *linear* in capital income. (Remember that the government is free to use an arbitrarily nonlinear system.) The optimal tax system *subsidizes* the capital income of surprisingly highly skilled people and *taxes* the capital income of surprisingly low-skilled people. While seemingly regressive, this tax system actually provides better social insurance. Intuitively, the tax system provides better incentives because it deters people from accumulating lots of wealth and then pretending to be low-skilled. These better incentives expand the scope for social insurance.

The heterogeneity in tax rates across people means that the Mirrlees prescription for optimal capital taxes differs from the Ramsey prescription for optimal capital tax rates. However, the two approaches do coincide in their recommendations for total and average capital income taxes. The Mirrlees approach recommends subsidies on some people and taxes on others. However, one can prove that, in the optimal tax system, both the average tax rate (across people) and the total tax revenue from capital income taxes are zero at every date. (See Kocherlakota 2006, for a survey article on the Mirrlees approach.)

Making Government Endogenous

In both the positive approach and the normative approach, the government is a preprogrammed robot during the life of the economy. It would be useful to develop models in which the government is another economic actor (or, even more realistically, a collection of economic actors) that makes choices at intervals based on its information. Such models would allow us to understand what forces lead to the kinds of policy choices that we see in reality. (See Persson and Tabellini 2000, for a much more complete discussion of these issues.)

These models need to capture at least two types of conflict. One source of conflict is heterogeneity. Households differ in their attributes and so in their preferences over policies. Old people have shorter horizons and typically prefer to set public investment to lower levels than young people. People with lots of capital prefer lower capital tax rates than do people with little capital. People with lots of nominal debt would like to raise nominal interest rates; their lenders prefer the opposite.

There is a great deal of research studying these kinds of conflicts. Unfortunately, it has been hard to generate the kind of robust answers that macroeconomists have obtained from the positive and normative approaches. There is no real consensus about how to model the games that get played by the different groups. Some researchers use voting games, while others use bargaining games. Some researchers treat conflicts in isolation, while others model conflicts as being resolved in bundles. These different modelling choices generate substantially different predictions about policy formation.

In a classic article, Kydland and Prescott (1977) set forth a second source of conflict. Suppose the world lasts two periods, and a government wants to raise taxes to finance purchases using capital income taxes and labour income taxes. Assume that all households are identical – so that the first type of conflict is removed – and that the government cares only about maximizing household welfare. It would seem that all sources of conflict have been removed in this situation.

But this is not true. The period 1 government's preferences over period 2 capital taxes are

fundamentally different from the period 2 government's preferences. In period 2, the amount of capital in the economy is fixed – there is no way to get any more. The period 2 government would like to set a high tax rate on this fixed tax base to raise as much revenue as possible.

In period 1, though, the amount of capital in period 2 has yet to be determined. The period 1 government has to consider how the tax rate in period 2 affects the size of the period 2 tax base. Its preferred period 2 tax rate is much smaller than the tax rate that the period 1 government likes.

Thus, even if governments at different dates are all benevolent, there is a dynamic conflict between them. How this conflict gets resolved is, again, a nontrivial matter. The dynamic game does have a unique equilibrium in a finite horizon. Unfortunately, this unique equilibrium is unrealistic in most countries: capital tax rates are set very high in every period. On the other hand, if the game has an infinite horizon, then there are an infinite number of equilibrium outcomes, including ones with high capital tax rates, low capital tax rates, and paths that vary between the two (see Chari and Kehoe 1990). The predictive power of the model is then quite limited.

Conclusions

There is an old joke to the effect that if you ask 10 macroeconomists about a policy question, you'll get 11 different answers. This joke provided a disturbingly accurate picture of the state of the field in the 1970s and 1980s. To a remarkable extent, it was no longer applicable as of 2005. There is a profession-wide consensus on methods that simply did not exist in the early 1980s. This consensus has led to a set of results about monetary and fiscal policy that are sharp, robust and surprising.

See Also

- ▶ Optimal Taxation
- ▶ 'Political Economy'
- ▶ Ricardian Equivalence Theorem
- ▶ Social Insurance

Acknowledgment I thank Barbara McCutcheon for her comments. The opinions expressed herein are mine and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Bibliography

Abel, A., N.G. Mankiw, L. Summers, and R. Zeckhauser. 1989. Assessing dynamic efficiency: Theory and evidence. *Review of Economic Studies* 56: 1–19.

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25: 380–396.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Cass, D., and K. Shell. 1983. Do sunspots matter? *Journal of Political Economy* 91: 193–227.

Chamley, C. 1986. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica* 54: 607–622.

Chari, V.V., and P. Kehoe. 1990. Sustainable plans. *Journal of Political Economy* 98: 783–802.

Chari, V.V., and P. Kehoe. 1999. Optimal fiscal and monetary policy. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford, Vol. 1C. Amsterdam: North- Holland.

Chari, V.V., L. Christiano, and P. Kehoe. 1996. Optimality of the Friedman rule in economies with distorting taxes. *Journal of Monetary Economics* 37: 203–223.

Correia, I., and P. Teles. 1999. The optimal inflation tax. *Review of Economic Dynamics* 2: 325–346.

Correia, I., Nicolini, J. -P. and Teles, P. 2004. *Optimal fiscal and monetary policy: Equivalence results*. Working paper, Centre for Economic Performance, London School of Economic.

Fernandes, A., and C. Phelan. 2000. A recursive formulation for repeated agency with history dependence. *Journal of Economic Theory* 91: 223–247.

Green, E. 1987. Lending and smoothing of uninsurable income. In *Contractual agreements for intertemporal trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Lucas, R.E. Jr. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit, and Banking* 12: 696–715.

Judd, K. 1985. Redistributive taxation in a perfect foresight model. *Journal of Public Economics* 28: 59–84.

Kocherlakota, N. 2006. Advances in dynamic optimal taxation. In *Advances in economics and econometrics: Theory and applications: Ninth World congress of the econometric society*, ed. R. Blundell, W.K. Newey, and T. Persson, Vol. 1. Cambridge: Cambridge University Press.

Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–491.
Lucas, R.E. Jr., and N. Stokey. 1987. Money and interest in a cash-in-advance economy. *Econometrica* 55: 491–513.
Obstfeld, M., and K. Rogoff. 1983. Speculative hyperinflations in maximizing models: Can we rule them out? *Journal of Political Economy* 91: 675–687.
Persson, T., and G. Tabellini. 2000. *Political economics, explaining economic policy*. Cambridge, MA: MIT Press.
Prescott, E. 2005. The transformation of macroeconomic policy and research. In *Les Prix Nobel. The Nobel Prizes 2004*, ed. T. Frängsmyr. Stockholm: Nobel Foundation Online. Available at http://nobelprize.org/nobel_prizes/economics/laureates/2004/prescott-lecture.pdf. Accessed 18 Oct 2006.
Sargent, T., and N. Wallace. 1975. Rational expectations, the optimal monetary instrument and the optimal money supply rule. *Journal of Political Economy* 83: 241–254.
Sargent, T., and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3): 1–18.
Woodford, M. 1994. Monetary policy and price level determinacy in a cash-in- advance economy. *Economic Theory* 4: 345–380.

Monetary Approach to the Balance of Payments

Mario I. Blejer and Jacob A. Frenkel

JEL Classifications
F3

The monetary approach to the balance of payments is an analytical formulation which emphasizes the interaction between the supply and the demand for money in determining the country's overall balance of payments position. It could be seen as an extension, to the case of an open economy, of traditional closed-economy monetary theory, which stresses the stability of the money demand function and considers the various channels through which changes in the money supply affect the economy. If changes in the money supply are not matched by equivalent changes in demand, then a stock disequilibrium arises. In responding to the stock disequilibrium, individuals alter their spending patterns. These adjustments are subject to the budget constraints which link the excess flow supply of money to the corresponding excess flow demand for goods and services. In a closed economy nominal income rises and interest rates may change so as to eliminate the disequilibrium in the money market; the increase in prices, and possibly output, in conjunction with the change in interest rates, raises the nominal demand for money to a level equivalent to the rise in the nominal money stock.

In contrast to the closed economy, the open economy has additional channels through which monetary imbalances are resolved. In the open economy changes in the money stock can arise from domestic credit creation as well as from the foreign exchange operations of the monetary authorities. As a result, the monetary approach to the balance of payments stresses that money market disequilibria are reflected not only in changes in nominal income but also in the country's overall balance of payments, as represented by changes in foreign exchange reserves. Thus the monetary approach to the balance of payments focuses on the relation among prices, output, interest rates, *and* the balance of payments.

In developing the simplest version of the monetary approach to the balance of payments, it is assumed that the country is small, fully employed, that it has a fixed exchange rate, and that there is perfect international mobility of goods and financial assets. These assumptions mean that domestic prices and interest rates equal their respective (exogenously given) world values, and that output is determined exogenously. Under such circumstances, any disequilibrium emerging from the money market is fully reflected in the balance of payments. For example, an excess supply of money arising from domestic credit expansion results in a loss of international reserves. This loss reduces the outstanding money stock to its equilibrium level consistent with the given demand. By concentrating on the direct connection between the money market and the balance of payments, rather than working through the implied changes in the goods or financial assets

markets, the monetary approach distinguishes itself from other analytical approaches to balance of payments theory.

The Development of the Approach

The monetary approach to the balance of payments has a long intellectual history originating with the 18th-century writings of David Hume. The continuity of its development, however, was reversed for upwards of a quarter of a century by the events of the 1930s. This included the international monetary collapse of 1931 and after, and the 'Keynesian revolution'.

The modern revival of the monetary approach originated with the writings of James Meade in the early 1950s followed by Harry G. Johnson and Robert A. Mundell in the 1960s. At the same time, important contributions to the formal development of the approach were carried out, under the leadership of Jacques J. Polak at the International Monetary Fund, thereby yielding analytical foundations to the Fund's operational practices.

By the late 1960s a long series of articles, subsequently collected in Mundell (1968, 1971), Frenkel and Johnson (1976) and International Monetary Fund (1977), gave an increasing stimulus to the rapid development of theoretical and empirical work on the monetary approach. Many of the contributions are surveyed in Kreinin and Officer (1978) and in Frenkel and Mussa (1985).

Theoretical Underpinnings

In order to assess the major implications of the monetary approach to the balance of payments, it is useful to present a simplified model which embodies the central characteristics of the analytical approach. The stripdown basic model considers a small, fully employed country operating under a fixed exchange rate system and assumes full integration of domestic and foreign goods and capital markets. Perfect arbitrage determines the prices of domestic commodities and of financial assets.

Because of its concentration on the money market, the monetary approach to the balance of payments involves the explicit specification of the money supply process and of a demand for money function. The supply of money (M^s) is the product of the stock of high-powered money (H) and the money supply multiplier (m) where the latter reflects the behaviour of asset-holders and the banking system:

$$M^s = mH. \tag{1}$$

By definition, the stock of high-powered money (the liabilities of the monetary authorities) is equal to the domestic currency value of the stock of international reserves, eR (where e is the exchange rate, defined as the domestic-currency price of foreign exchange, and R is the foreign currency value of international reserves), and the domestic asset (net of liabilities) holdings of the monetary authorities (D):

$$H = eR + D. \tag{2}$$

The demand for real money balances is specified as a positive function of real income and a negative function of the opportunity cost of holding money. This opportunity cost is measured by the yield on alternative financial assets, usually represented by the rate of interest. The demand for money in nominal terms (M^d) can be written as:

$$M^d = Pf(Y, i) \tag{3}$$

where P denotes the domestic price level, Y is the level of domestic real income, and i stands for the domestic nominal interest rate.

Money market equilibrium implies that $M^s = M^d$. Under the assumptions of the simplified model, the mechanism responsible for maintaining equilibrium operates through changes in international reserves. Accordingly, using Eqs. (1), (2), and (3) the (endogeneously determined) stock of international reserves can be specified as:

$$R = g(P, Y, i, m, D). \tag{4}$$

Equation (4) represents the key relationship implied by the monetary approach to the balance of payments under a fixed exchange rate system. The assumed specifications of M^s and M^d imply that an increase in real income and in (world) prices raises the stock of international reserves while an increase in the rate of interest, in the money multiplier, and in the net domestic assets of the central bank reduces the stock of international reserves. These changes in the stock of international reserves are reflected in balance of payments surpluses or deficits. The size of the income and interest rate effects depends on the elasticities of the money demand function. In this simple model a rise in the money supply brought about by an open-market purchase (an increase in D) is completely offset by a corresponding fall in R.

An important implication of the analysis is that under a fixed exchange rate regime the nominal money supply is no longer within the direct control of the monetary authorities and becomes an endogenous variable of the system. The monetary authorities, however, do retain control over the volume of domestic credit, which is one of the sources of money creation. The distinction between high- powered money and its domestic credit component becomes crucial: the central bank controls the latter but not the former. Given a rate of growth in the demand for money, an equivalent growth in the supply can be obtained by an appropriate increase in domestic credit. However, if the rate of domestic credit expansion differs from the growth in demand, then the difference between the two is made up by changes in net foreign assets, brought about through a balance of payments surplus or deficit.

Extensions and Analytical Applications

The simplified model presented in the previous section may be seen as a prototype of the monetary approach to the balance of payments and can also be regarded as a representation of its long-run equilibrium characteristics, when all adjustments have taken place. In these circumstances, monetary imbalances tend to affect primarily the balance of payments. However, if the degree of international capital mobility is not high and if the share of non-tradeable goods in GNP is relatively high, then the speed of adjustment to monetary disturbances is reduced. In the short run, therefore, monetary imbalances also affect prices, output, and interest rates, and the relative importance of these effects depends on various factors such as the nature of exchange rate management, the degree of openness of the economy in both the goods and the capital markets, the proportion of tradeable and non-tradeable goods, the degree of resource utilization, the degree of nominal and real wage rigidities, and so forth. Many of those elements have been specifically modelled within the framework of the monetary approach, and the effects of considering different sets of alternative assumptions have been carefully analysed. A central feature of most of the short-term extensions of the basic model is that the excess demand in the commodity market, caused by excess supply in the money market, results in a combination of price increases (which reduce the real value of the outstanding nominal money stock) and balance of payments deficits (which, by depleting the level of international reserves, reduce the level of the nominal money stock). These changes take place in addition to income changes, which in the short run depend on the degree of resource utilization and on the degree to which the public has anticipated the monetary expansion. The effects of monetary disequilibrium fall more heavily on the domestic price level and on the domestic interest rate, and less on the balance of payments, the lower the degree to which the economy is integrated into the world markets for goods and capital. Therefore, the effects of monetary imbalances on the domestic price level and interest rate are stronger the larger are the relative shares of nontraded goods and financial assets, and the more prohibitive are import tariffs, quantitative restrictions and exchange controls.

Further extensions of the basic framework have considered the effects of exchange rate changes on prices and on the balance of payments. In contrast with other approaches to balance of payments analysis (notably the elasticities approach), the monetary approach stresses that

the effects of a once-and-for-all exchange rate adjustment in a small economy are transitory. A devaluation (a rise in e) raises the price of internationally tradeable goods. This increase in price reduces the real value of the nominal money stock and, in order to restore money market equilibrium, a balance of payments surplus is generated as foreign exchange reserves flow into the country. As monetary equilibrium is restored, the flow of reserves stops.

The negative relationship between the rate of expansion of domestic credit and the rate of change of foreign exchange reserves implied by the monetary approach does not necessarily imply a unidirectional causality. In fact, it is possible that central banks manipulate their domestic assets in order to sterilize the impact of exogenous changes in foreign reserves on the domestic supply of money. Assume, for example, a reserve-gaining country which desires to avoid an increase in its money supply. The central bank will tend to counteract the inflow of reserves by reducing its credit to commercial banks or its lending to the government. The required volume and the effects of these sterilization operations could easily be analysed within the framework of the monetary approach, if the parameters underlying the money demand function and the money supply process were known.

The effects of income growth and of external shocks can also be examined within the same set-up. As shown by Eq. (4), changes in the level of income have a direct impact on the balance of payments through their effect on the demand for money. Therefore, an acceleration in a country's rate of growth, by increasing the demand for liquidity, tends to improve the balance of payments provided that domestic credit policy does not expand accordingly. Similarly, external shocks, such as terms of trade changes, which affect domestic activity, also affect the balance of payments through the same mechanism. In particular, a negative external shock which reduces real income results in a once-and-for-all reduction in the demand for money and (in the absence of domestic credit policy) results in foreign exchange reserves.

The basic model can also be used to determine the effects of commercial policies such as an import tariff. A tariff affects the balance of payments by raising the domestic price level and thereby, by lowering the real value of the outstanding money stock. These changes are likely to induce an excess demand for money which, other things being equal, results in an inflow of international reserves. Similar principles can be used to analyse the effects of other forms of taxation and commercial policies.

Finally, the model could be generalized to the 'large-country' case. When the country is not small relative to the rest of the world, one needs to take account of the impacts of its policy and economic behaviour on the world price of tradeable goods and on the world rate of interest. While the monetary mechanism of balance of payments adjustment is more complicated for the large-country case, the basic elements of this mechanism are essentially the same. Starting from a situation in which the domestic nominal money supply is below its long-run equilibrium level and, correspondingly, the foreign money supply is above its long-run equilibrium level, reserve flows associated with trade imbalances gradually move the economic system to long-run equilibrium by raising the domestic money supply and reducing the foreign money supply to their respective long-run equilibrium levels. As in the case of the small country, the essential ingredient underlying this adjustment process is the relationship through which a deficiency in a country's money supply relative to its long-run equilibrium level leads to an excess of domestic income over domestic expenditure which implies a trade surplus which brings an inflow of foreign exchange reserves and a gradual restoration of money balances to their long- run equilibrium level.

In the two-country world, it remains true that a given initial divergence of a country's money supply will ultimately lead to a cumulative payments surplus and change in reserves just equal to this initial divergence, assuming there is no change in the non-reserve assets of central banks.

Overview

In general, a proper analysis of the balance of payments emphasizes the budget constraint

imposed on the country's international spending and views the various accounts of the balance of payments as the 'windows' to the outside world, through which the excesses of domestic flow demands over domestic flow supplies, and of excess domestic flow supplies over domestic flow demands, are cleared. Accordingly, surpluses in the trade account and the capital account, respectively, represent excess flow supplies of goods and securities, and a surplus in the money account reflects an excess domestic flow demand for money. Consequently, in analysing the money account, or more familiarly the rate of increase or decrease in the country's international reserves, the monetary approach focuses on the determinants of the excess domestic flow demand for or supply of money.

Although it concentrates on the money account of the balance of payments, the monetary approach should, in principle, give an answer not different from that provided by a correct analysis in terms of the other balance of payments accounts. The surplus or deficit in the goods account (more generally the current account) measures the extent to which the economy's income is greater than consumption ('absorption') and the economy is therefore accumulating claims on future income (assets) from abroad or vice versa. By virtue of the budget constraint, the sum of the deficit on the capital account (net purchase of foreign securities) and the surplus on the money account equally represents the accumulation of foreign assets (decumulation if negative). The so-called 'absorption approach' to the balance of payments, associated with Sidney Alexander (1952), emphasizes the rate of accumulation or decumulation of foreign assets (securities plus money). In so doing, it differs from the 'elasticity approach', which emphasizes relative-price mechanisms.

The monetary approach selects for emphasis a subset of the spectrum of foreign assets whose accumulation or decumulation is emphasized by the absorption approach. The main reasons for this are, firstly, that the accumulation of foreign assets does not necessarily imply the accumulation of money through the balance of payments – it may mean the opposite, as for example when a monetary policy of lowering interest rates leads domestic asset-holders to move their funds from domestic to foreign securities. Secondly, the monetary authorities, in their role as stabilizers of the exchange rate in a fixed rate system, are concerned with what causes the stock of international reserves to change and how to prevent such changes. Thirdly, the monetary authorities, as the ultimate source of domestic money, control the rate of change of the domestic credit component of the monetary base –the other component being international reserves. The assumption that the residents of the country have a demand for money which depends on variables at least in part different from those that determine the quantity of domestic credit extended by the banking system, or alternatively, that the rate of change of money demanded (the rate of hoarding) is independent of the rate of change of the domestic credit source component of the monetary base, implies that the money account of the balance of payments is influenced directly by monetary policy.

See Also

- ▶ Absorption Approach to the Balance of Payments
- ▶ Elasticities Approach to the Balance of Payments
- ▶ International Finance
- ▶ Purchasing Power Parity
- ▶ Specie-Flow Mechanism

Bibliography

Alexander, S.S. 1952. Effects of a devaluation on a trade balance. *IMF Staff Papers* 2: 263–278.

Frenkel, J.A., and H.G. Johnson, eds. 1976. *The monetary approach to the balance of payments*. London/Toronto: Allen & Unwin/University of Toronto Press.

Frenkel, J.A., and M.L. Mussa. 1985. Asset markets, exchange rates and the balance of payments. In *Handbook of international economics*, ed. R.W. Jones and P.B. Kenen, vol. II. New York: Elsevier.

International Monetary Fund. 1977. *The monetary approach to the balance of payments*. Washington, DC: International Monetary Fund.

Kreinin, M., and L. Officer. 1978. *The monetary approach to the balance of payments: A survey,* Princeton Studies in International Finance No. 43. Princeton: Princeton University Press.

Meade, J.E. 1951. *The theory of international economic policy, vol. I. The balance of payments*. Oxford: Oxford University Press.
Mundell, R.A. 1968. *International economics*. New York: Macmillan.
Mundell, R.A. 1971. *Monetary theory.* Pacific Palisades: Goodyear Publishing Company.

Monetary Base

Charles Goodhart

A key characteristic of bank deposits is that they carry a guarantee of convertibility at sight, or after due notice, into cash. In order to maintain such convertibility, a bank needs to hold reserves of cash. Historically such cash mostly took the form of metallic coin, that is, gold, silver or copper. Nowadays the cash base mostly consists of the liabilities of the Central Bank, primarily notes, but also bankers' balances at the Central Bank which the bankers can, if they wish, withdraw in note form to add to their own cash holdings. The monetary base, mostly consisting of Central Bank notes in the hands of the public and in the tills of the banks, is so called because it provides the cash base on which the much larger superstructure of convertible deposits is erected.

Such cash holdings have generally not paid any interest; indeed it would be difficult to devise a technical method of paying interest on notes and coins. Accordingly commercial banks have had an incentive to economize on their holdings of such zero-yielding cash assets, restricting them to the minimum required in order to satisfy the convertibility requirements, while protecting themselves against large-scale deposit withdrawals by holding a range of liquid (near-cash) assets, which could be rapidly transformed into cash (liquified) at short notice, but which nevertheless offered a reasonable yield. The main component of such second-line liquidity has historically been short-term commercial or Treasury bills; the liquidity of such bills has, in turn, been enhanced by the willingness of the Central Bank always to re-discount, and to maintain a market in, such bills, though on occasions at a penalty price.

Although most banking panics have actually been caused by growing concern about some banking institutions' solvency, the actual event that forces closure in such cases, at any rate in the earlier years of banking, was the inability of the bank to continue paying out its depositors, when the bank was faced with a 'run' of deposit withdrawals, in cash. Accordingly, to show how strong the bank was, banks tended to windowdress their balance sheets on publication dates, with more cash and liquid assets than they in reality normally held. Meanwhile, the monetary authorities, noting that the proximate cause of failure was a shortage of cash reserves, often imposed requirements, whether backed by legal force or though moral suasion, that the banks should hold a certain percentage of their assets in cash form, and, in some cases, an additional percentage in some specified set of liquid assets. This latter policy had certain inherent deficiencies. First, to the extent that such balances were actually *required* to be held, they could not then be legally used to meet withdrawals, so the effective available reserves became the margin of 'free reserves' in excess of requirements. Moreover, any, even temporary, decline of reserve holdings below the required level was taken as a signal of weakness and distress in itself. Second, the requirements for banks to hold larger zero-yielding and low-yielding balances, then they would have voluntarily done, adversely affected their profitability. This not only weakened their ability to compete, but in some cases may have encouraged the banks to undertake a riskier strategy in order to restore their profitability, thereby negating the initial intentions of the authorities.

Still, the banks *had* to hold a certain proportion of cash reserves, in order to remain in business: indeed there was often a certain observed regularity and stability in the ratio of their cash reserves to deposits – though this was sometimes the consequence either of windowdressing or of official requirements. Such stability in the banks' reserve deposit ratio, combined with the fact that the cash base largely represented the liabilities of the Central Bank, led economists to construct a theory

of the determination, and control, of the money stock. This is based on certain simple accounting identities. The money stock (M) is defined as comprising two main components, being respectively currency (C) and bank deposits (D) held by the general public. It is, therefore, possible to set down the identity

$$M = D + C$$

which must hold exactly by definition. Similarly it is possible to define the sum of currency held by the general public (C) and the cash reserves of the banking sector (R) as 'high-powered money' or 'monetary base' (H). Again, the additional identity can be formed

$$H = R + C$$

By algebraic manipulation of these two identities it is possible to arrive at a third identity

$$M = H(1 + C/D)/(R/D + C/D)$$

describing the money stock in terms of the level of high-powered money and two ratios, *R/D,* the banks' reserve/deposit ratio, and *C/D,* the general public's currency/deposit ratio. Since this relationship is also an identity, it always holds true by definition; changes in the money stock can therefore be expressed in terms of these three variables alone. To be able to express changes in the money stock in terms of only three variables has considerable advantages of brevity and simplicity. Nevertheless, the use of such an identity does not in any sense provide a behavioural theory of the determination of the stock of money.

The associated behavioural story rests upon a supposed 'multiplier' process, the monetary base multiplier. On this thesis, the Central Bank undertakes open-market operations, in order to vary its own liabilities, and, in the process, the reserve base of the banking system. When, for example, the Central Bank sells an asset, the purchaser, probably a non-bank, pays for the purchase by a cheque on her own bank, so that bank's balance with the Central Bank is reduced. Then, on this story, that bank sells an asset itself in order to restore its depleted cash balance. This second purchaser, again probably a non-bank, paying again by cheque, will by so doing transfer cash reserves from his own bank to the first bank in order to pay for the purchase, but in the process the cash shortage will be transferred to this second bank. And so the multiplier process will continue. So long as the Central Bank does not re-enter the market in order to buy assets, its initial open market sale will cause a multiple fall in bank assets and deposits, the size of which multiplier will depend on the *C/D* and the *R/D* ratios described above.

In practice, however, the banking system has virtually *never* worked in that manner. Central Banks have, indeed, made use of their monopoly control over access to cash and their power to enforce that by open market operations, but for the purpose of making effective a desired level of (short-term) interest rates, not to achieve a pre-determined quantity of monetary base or of some monetary aggregate. The various influences, external forces and objectives that have affected the authorities' views of the appropriate level of interest rates have varied over time, including such considerations as a desire to maintain a fixed exchange rate, for example on the Gold Standard, or to encourage investment, or, more recently, to influence the pace of monetary growth itself.

Indeed, Central Banks have historically been at some pains to assure the banking system that the institutional structure is such that the system as a whole can *always* obtain access to whatever cash the system may require in order to meet its needs, though at a price of the Central Bank's choosing: and there has been a further, implicit corollary that that interest rate will not be varied capriciously. The whole structure of the monetary system has evolved on this latter basis, that is, that the untrammelled force of the monetary base multiplier will *never* be unleashed. Furthermore, recent institutional developments, notably the growth of the wholesale inter-bank liability market, imply that the monetary base multiplier no longer would, or could, work in the textbook fashion. The development of liability management, through such wholesale markets, means that

commercial banks now respond to a loss of liquidity, whether from a Central Bank open-market sale or some other source, by bidding for additional funds in such liability markets, rather than by selling assets. In these circumstances, a loss of cash reserves to the banking system, driving these below an acceptable minimum, will simply have the effect of driving interest rates upwards, both in liability markets, and more generally on deposit and asset rates, without having initially any direct effect on monetary quantities. In the short run, then, interest rates can rise, in a virtually limitless spiral until extra reserves are attracted into the system, whether from the Central Bank, or elsewhere. In the somewhat longer run, however, the rise in interest rates will subsequently bring about a reallocation by both bank depositors and bank borrowers of their funds, which will, in general, have the effect of bringing about a transfer of funds from non-interest-bearing narrow monetary aggregates, and also leading to a reduction of now more expensive bank borrowing.

In short, the behavioural process runs from an initial change in interest rates, whether administered by a Central Bank or determined by market forces, to a subsequent readjustment in monetary aggregate quantities: the process does *not* run from a change in the monetary base, working via the monetary base multiplier, to a change in monetary aggregates, and thence only at the end of the road to a readjustment of interest rates. In reality, the more exogenous, or policy-determined, variable is the change in (short-term) interest rates, while both the monetary base and monetary aggregates are endogenous variables. This reality is, unfortunately, sharply in contrast with the theoretical basis both of many economists' models, and also of their teaching. The fact that it is commonplace to find economists treating the monetary base and/or the money stock as exogenously determined in their models does not mitigate the error; the fact is that this approach is simply incorrect. Moreover, when it comes to a practical, historical account of how Central Banks have *actually* behaved, most economists, even including those who treat the money stock as exogenously determined in their own theoretical models, accept the reality that Central Banks have generally sought to set interest rates, according to various objectives, and that the monetary base and money stock has, therefore, been endogenously determined. The argument then switches from an analysis of how the money stock is determined, to the normative question of whether the present techniques of monetary control adopted by most Central Banks are appropriate and reasonable, or whether the Central Banks should, instead, adopt monetary base control, and thenceforth actually seek to operate the kind of control technique, which is to be found in textbook and theory, but very rarely operated in practice.

The arguments against Central Bank discretionary control of interest rates are several. First, that the authorities do not have sufficient understanding to be able to adjust interest rates in a stabilizing fashion. The second, is that the authorities would be under (political) pressures to hold interest rates down, since rising interest rates are politically unpopular. Such pressures could cause interest rates to be adjusted too little and too late, with the consequence that monetary growth would have, and indeed can be shown to have empirically, a pro-cyclical bias, so that monetary policy would act in a de-stabilizing fashion.

The positive argument for monetary base control is that this would provide a clear and accountable guide for Central Banks. It would remove the political element in the determination of interest rates and give market forces a greater role in setting this key price. Moreover, the medium and longer term stability of the relationship between monetary growth and nominal incomes would allow adherence to closer monetary control, via operating through the monetary base multiplier, to result in greater long term stability of nominal incomes and inflation.

In recent years, many Central Banks have accepted, in part, the argument that (political) pressures have led to some bias to delay, and have adopted publicly announced monetary targets as a main intermediate objective for their policies. They maintain, however, that structural changes and other unforeseeable forces can change the relationship between any monetary aggregate and nominal incomes quite markedly,

even over short periods, so that a degree of discretion in maintaining monetary control remains essential. Furthermore, and more closely related to the question of monetary base control, they believe that their present techniques, mostly involving direct interest rate adjustments, remain sufficient to the task.

In particular Central Banks assert that, given the present institutional structure, the attempt to enforce and impose a certain predetermined level of monetary base on the banking system, irrespective of that system's requirements at the time for cash reserves, would lead to a devastating increase in the volatility of interest rates. Moreover, with the resulting effect on the monetary aggregates occurring after a lag, which could be quite long as individual agents adjusted to the rapidly changing level of interest rates, the ultimate effect of the initial shock would itself be unpredictable, and not necessarily desirable. In this context, the experience of the Federal Reserve in the US, which in October 1979 adopted a moderated version of monetary base control, whose potential extreme effects were alleviated by allowing the system access to the discount window, is instructive. The volatility of short-term interest rates increased four-fold during the period of the experiment, lasting from October 1979 until September 1982; moreover this volatility also passed through into the long-term bond market and the foreign exchange market. Despite trying to control the reserve base of the monetary system, the exercise resulted in even greater short-term volatility in the rate of growth of the targeted monetary aggregate, M1. The result, therefore, was much greater market volatility, without any particular success in achieving a more stable path for the monetary aggregates.

Proponents of the switch to monetary base control often accept that the present institutional structure is, indeed, geared to present Central Bank operating techniques, and would, perhaps, be likely to suffer greater interest rate volatility, were monetary base control methods to be adopted. But they then claim that the commitment to, and experience with, monetary base control methods would lead the institutional structure to adapt reasonably quickly so as to moderate such interest rate volatility. There was little sign of that occurring in the United States by 1982. Be that as it may, the opponents of monetary base control argue that those same institutional changes would, inter alia, probably lead institutions to hold larger cash reserve balances on average, but be prepared to allow these to adjust much more in response to the authorities' actions to change the cash base. If so, the new institutional structure would cause changes that would not only lead to much greater variability in the *R/D* ratio, which would itself lessen the reliability and predictability of the money multiplier, but could also well lead to disintermediation to other financial intermediaries, might be better placed to protect themselves from the instability to the system caused by the authorities' actions. Under such circumstances, therefore, the advantages that are now posited for monetary base control on the basis of calculations of the money multiplier constructed in the present institutional setting, might well erode, if not vanish entirely, should the policy regime actually change.

See Also

- ▶ High-Powered Money and the Monetary Base
- ▶ Monetary Policy
- ▶ Money Supply
- ▶ Quantity Theory of Money

Bibliography

Classical

Keynes, J.M. 1930. *A treatise on money*. London: Macmillan.

Phillips, C.A. 1920. *Bank credit*. New York: Macmillan.

Historical

Cagan, P. 1965. *Determinants and effects of changes in the stock of money, 1875–1960*. New York: Columbia University Press for the National Bureau of Economic Research.

Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Frowen, S.F., et al. 1977. *Monetary policy and economic activity in West Germany*. Stuttgart/New York: Gustav Fischer Verlag.

Contemporary
Aschheim, J. 1961. *Techniques of monetary control*. Baltimore: Johns Hopkins Press.
Bank for International Settlements. 1980. *The monetary base approach to monetary control*. Basle: BIS.
Burger, A.E. 1971. *The money supply process*. Belmont: Wadsworth.
Dudler, H.-J. 1984. *Geldpolitik und ihre theoretischen Grundlagen*. Frankfurt: Fritz Knapp Verlag.
Federal Reserve Bank of Boston Conference Series. 1972. *Controlling monetary aggregates. II: The implementation.* Boston: FRB.
Federal Reserve Staff Studies. 1981. *The new monetary control procedures*. Washington, DC: Board of Governors of the Federal Reserve System.
Goodhart, C.A.E. 1984. *Monetary theory and practice*. London: Macmillan.
H.M. Treasury and The Bank of England. 1984. *Monetary control.* Cmd 7858. London: HMSO.
Meigs, A.J. 1962. *Free reserves and the money supply.* Chicago: University of Chicago Press.
Tobin, J. 1963. Commercial banks as creators of 'money'. In *Banking and monetary studies*, ed. D. Carson. Homewood: Richard, D. Irwin.

Monetary Business Cycle Models (Sticky Prices and Wages)

Christopher J. Erceg

Abstract
Monetary business cycle (MBC) models are general equilibrium models designed to analyse how monetary shocks affect output, prices, and interest rates. This article describes the analytic framework underlying sticky prices and wages in modern MBC models, and highlights the prominent role that these rigidities play in the transmission of nominal and real shocks.

Keywords
Cost-push inflation; Friedman, M; Inflation; Inflationary expectations; Intertemporal optimization; Keynes, J. M; Lucas, R; Microfoundations; Monetary business cycle models; Monetary policy rules; Monetary shocks; Monetary transmission mechanism; New Keynesian Phillips curve; Nominal rigidities; Nominal shocks; Nominal wage inflation; Output gap; Phelps, E; Phillips curve; Price dynamics; Rational expectations; Real business cycles; Real rigidities; Real shocks; Staggered contracts model; Sticky prices; Sticky wages; Technology shocks; Unemployment

JEL Classification
D4; D10

Since the earliest analysis of the monetary transmission mechanism by pre-eminent classical economists of the 18th and early nineteenth century, sticky prices and wages have been identified as playing a central role (Humphrey 2004). The classical economists believed that prices adjusted gradually to a change in the nominal money stock, so that monetary changes could exert substantial short-run effects on output. Nominal wages were regarded as particularly slow to change, and thus helped account for gradual price adjustment by mitigating short-run pressures on factor costs.

The classical economists and their successors used this framework both to guide recommendations about policy and to evaluate alternative monetary regimes. For example, the belief that prices would respond slowly to a monetary contraction led Thornton and Ricardo to recommend a gradualist approach to deflation.

Early Keynesian Models, and Some Critiques

A major contribution of Keynes (1936) and prominent successors such as Hicks to understanding the monetary transmission mechanism consisted in developing an explicit theoretical framework expressed in terms of equilibrium conditions in goods and asset markets. This IS–LM framework was of great value in illuminating the channels through which monetary shocks affected interest rates and output. However, the assumption of fixed prices and wages was a major shortcoming. It was eventually supplanted by the famous 'Phillips curve' relation linking nominal wage

inflation to the unemployment rate, or variants relating price inflation to the output gap:

$$p(t) - p(t-1) = b * (y(t) - y(t)*) b > 0 \quad (1)$$

where $p(t)$ is (the log of) the price level, $y(t)$ output, $y(t)$* potential output, and b is a parameter. The Phillips curve filled a missing link in earlier 'fixed price' IS–LM analysis by making it feasible to trace the dynamic effects of a monetary shock on prices and output. Thus, an initial rise in output following a monetary expansion boosts prices via (1), which in turn causes real balances and output to revert gradually to pre-shock levels. However, the Phillips curve had weak theoretical underpinnings, so that there was little economic rationale for what determined the sensitivity of prices to the output gap (that is, 'b' in (1)), for the activity variable(s) driving price dynamics, and for how inflation might be influenced by expectations.

A series of remarkable critiques beginning with the analysis of Friedman (1968) and Phelps (1968) provided impetus for developing more theoretically coherent models of price and wage dynamics. These authors argued that the Phillips curve should be augmented so that actual inflation depended directly on inflation expectations in addition to real activity. In this framework, output could be pushed above potential only through surprising private agents by keeping inflation above the level that they had forecast in previous periods. Since such surprises could not continue indefinitely, there could be no long-run trade-off between inflation and output: expansionary monetary policy would eventually raise expected inflation, resulting in higher inflation with no output stimulus.

Shortly thereafter, Lucas (1972) derived an 'expectations-augmented' Phillips curve in a clearly specified rational expectations model. Lucas adopted a signal extraction framework in which agents partly misinterpreted aggregate nominal shocks as shocks to the relative price of their own output good (due to limited information), and responded by adjusting their supply. Consistent with Friedman and Phelps, Lucas's model implied that aggregate output varied positively with the unanticipated component of inflation (with anticipated inflation exerting no real effects). But because unanticipated inflation was linked explicitly to a 'rational expectations' forecast error in Lucas's model – which would be expected to die away quickly as agents learned about the nature of underlying shocks – monetary shocks could exert only transient effects on output. This posed a serious challenge to traditional Keynesian models by suggesting that their ability to derive persistent effects in response to a monetary injection relied on ad hoc assumptions about price dynamics or expectations formation. Moreover, because only unanticipated changes in inflation affected output, Lucas's supply relation implied that any predictable policy was as good as any other (the 'policy ineffectiveness' proposition). This point, emphasized by Sargent and Wallace (1975), contrasted sharply with the activist policy stance that emerged from typical Keynesian models.

Monetary Transmission in Optimization-Based MBC Models

Since the mid-1990s a new generation of optimization-based MBC models has emerged that can generate 'traditional' Keynesian implications, but in a framework consistent with rational expectations and rigorous microfoundations. Roughly speaking, these new MBC models graft features that can induce sluggish price and/or wage adjustment onto an underlying real business cycle (RBC) model. (Blanchard 2000, and Taylor 1999, provide comprehensive surveys of the foundations of modern optimization-based MBC models, which were laid in a series of important contributions spanning several decades.)

To highlight salient features of the modern approach, it is helpful to examine a specific characterization of price-setting that has been utilized extensively in the literature. This relation, often called the 'New Keynesian Phillips curve (NKPC)', takes the form

$$p(t) - p(t-1) = B * E(t)[p(t+1) - p(t)] + b * (y(t) - y(t)*) \quad (2)$$

where $E(t)$ is the conditional expectation operator, and B is the discount factor.

Following Calvo (1983) and Yun (1996), the NKPC can be derived in a framework consistent with intertemporal optimization. Firms are assumed to behave as monopolistic competitors in the output market, and face downward-sloping demand curves for their distinctive products. Firms face a dynamic decision problem, because they are constrained to set a price that remains fixed in nominal terms over some random duration of time (referred to as the 'contract period', since firms are assumed to meet all demand at this fixed price until allowed to adjust). When a firm receives a signal enabling it to adjust its price, the firm resets it based on estimates of current and future marginal costs expected to prevail over the contract period. Because not all firms can change their price in a given period, price-setting is staggered – similar to the decentralized price-setting in actual economies. (See sticky wages and staggered wage setting for a discussion of the staggered contracts model.)

From a qualitative perspective, an MBC model in which prices are determined by the NKPC provides a conventional Keynesian account of the monetary transmission mechanism. Thus, a monetary shock increases nominal spending and, since the price level adjusts gradually, real output exhibits a persistent increase (in contrast to the transient real effects in Lucas's model). But as time passes, a larger proportion of firms receive a signal that allows them to raise their price in response to higher projected marginal costs. At an aggregate level, these relative price adjustments translate into a higher price level, which eventually restores real balances and output to pre-shock levels.

A major virtue of the microfounded approach is that it illuminates how the monetary policy rule and various structural features of the economy affect the transmission of nominal (and real) shocks. First, given that price adjustment is influenced directly by inflation expectations (as in (2)), monetary surprises have smaller effects on current inflation to the extent that the policy rule is expected to keep future inflation near target (that is, 'anchors' inflation expectations). Second, while the sensitivity of price inflation to the output gap ('b') clearly plays a key role in determining how quickly prices and output adjust to a monetary injection, this parameter is itself determined by features of the microeconomic environment. Quite intuitively, the parameter 'b' varies inversely with the mean duration of price contracts, so that longer contracts imply slower price adjustment and more persistent effects on output. But 'b' also depends on the responsiveness of firm-level marginal costs to the aggregate output gap, which in turn hinges on features of the specific microeconomic environment, including assumptions about factor mobility, capital utilization, and preferences. While some assumptions constrain 'b' to be large, a considerable literature has emerged showing how various 'real rigidities' such as firm-specific capital and labour can account for a low 'b' (even with fairly shortlived contracts); an insightful overview is provided in Woodford (2003). Such real rigidities appear important in allowing macro models to account for persistent output effects, while remaining consistent with disaggregate price data suggesting that firms change prices frequently (Bils and Klenow 2004).

The NKPC in (2), in which the output gap enters as the activity variable, is derived under the assumption that wages are fully flexible. But, as noted above, there is a long precedent in macroeconomics suggesting that sticky wages play an important role in the transmission process. As shown by Erceg et al. (2000), wage rigidity may be modelled in a framework isomorphic to that rationalizing price rigidity, with households acting as monopolistic suppliers of differentiated labour services. Christiano et al. (2005) have shown that a model that incorporates both wage and price rigidity can account remarkably well for the estimated dynamic effects of a monetary shock on output, prices, and interest rates. The presence of wage rigidity damps the rise in marginal cost due to a positive monetary injection, helping account for estimated persistence in the response of output. Moreover, a model including both types of rigidities can help account for the observed acyclicality of the real wage. By contrast, sticky prices alone imply too much procyclicality in the

real wage, while sticky wages alone (in the spirit of the classical economists and Keynes) imply too much counter-cyclicality.

Real Shocks and Alternative Policies in MBC Models

Given that monetary policy is widely perceived to have been much more stable since the mid-1980s, the literature has focused greater attention on how policy should respond to real shocks. Modern optimization-based MBC models are useful in this regard, because they provide a coherent framework for examining the transmission of real shocks in the presence of sticky wages and prices, and for assessing the role of monetary policy in affecting the economy's responses.

The presence of nominal rigidities can markedly affect the economy's responses to real shocks. Following Gali (1999), this can be illustrated by contrasting the effects of a persistent rise in technology in an RBC model (in which prices and wages are flexible) with the effects in an MBC model in which prices adjust according to Eq. (2). For simplicity, it is assumed that money demand takes the interest-inelastic form $M = P * Y$, and that the monetary authority holds the nominal money stock constant. In either model, money market equilibrium implies that output can expand only if prices fall proportionally. But as prices can drop instantaneously in the RBC model, the money supply rule is irrelevant in determining the real effects of the shock. Thus, the technology shock immediately boosts employment (as the substitution effect dominates the income effect), and the (percentage) jump in output exceeds the magnitude of the shock. By contrast, prices fall gradually in the MBC model, so that output is constrained to rise slowly given the fixed money stock. With prices determined by the NKPC, negative output gaps are required to induce prices to fall, consistent with employment remaining persistently below its pre-shock level.

As in the case of nominal shocks, the effects of real shocks may be highly sensitive to underlying features of the microeconomic framework, including those that determine the speed of price or wage adjustment. Thus, features that affect 'b' in the NKPC can markedly change how real shocks impact the economy. In the case of the technology shock, additional price sluggishness would translate into a smaller short-run expansion in output and greater employment contraction. Similarly, the inclusion of wage stickiness can markedly affect the responses to technology shocks. For example, while the NKPC derived under the assumption of flexible wages (Eq. (2)) implies that price inflation stabilization also keeps output at potential, the same policy could generate large output gap fluctuations if wages were sticky as well as prices.

Modern MBC models have also been applied fruitfully to normative issues. Optimal policy is derived by maximizing an objective function subject to the model's behavioural equations. Importantly, the objective function used in ranking alternative policies is typically derived from the utility functions of the economy's households (Woodford provides an extensive treatment).

A compelling message of this normative literature is that a well-designed policy must take account of its ability to influence inflation through an expectations channel. Thus, a policymaker acting 'under discretion' in an environment where inflation was determined by (2) would act as if the only margin on which to trade in devising a policy involved current inflation and output. However, such a 'discretionary' policy is suboptimal, because it fails to take account of its influence on the expected inflation term in (2). The analysis of Clarida et al. (1999) and Woodford shows that rules that are devised to take account of their influence on future expected inflation can perform much better in maximizing social welfare than discretionary policies that take future inflation as outside the central bank's control. For example, these authors show that well-designed policies can reduce substantially the impact of an adverse cost-push shock on current inflation (relative to the effects under discretion) by creating the perception that future policy will bring inflation back quickly to baseline.

Woodford emphasizes that the optimal monetary policy rule in an environment with forward-looking price-setting exhibits history dependence,

so that current monetary policy actions depend on past inflation and activity. This inertial character reflects that the optimal policy rule is derived in a framework in which future policy is expected to take full account of its influence on inflation expectations at earlier dates, much as optimal tax rules recognize their impact on previous investment decisions. Consistent with this history dependence, Woodford shows that it is generally optimal for monetary policy to reverse spikes in inflation above its target value, rather than follow the conventional wisdom of allowing 'bygones to be bygones'. Interestingly, this analysis provides strong support for some form of price level targeting – as recommended by Fisher and Keynes nearly a century ago – with the twist that the modern justification highlights the role it can play in optimally anchoring inflation expectations.

See Also

- ▶ Is–Lm in Modern Macro
- ▶ Monetary Transmission Mechanism
- ▶ Phillips Curve (New Views)
- ▶ Real Rigidities
- ▶ Sticky Wages and Staggered Wage Setting

Bibliography

Bils, M., and P. Klenow. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112: 947–985.

Blanchard, O. 2000. What do we know about macroeconomics that Fisher and Wicksell did not? *Quarterly Journal of Economics* 115: 1375–1409.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. Nominal rigidities and the dynamic effects of shocks to monetary policy. *Journal of Political Economy* 113: 1–45.

Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy: A new Keynesian perspective. *Journal of Economic Literature* 37: 1661–1707.

Erceg, C., D. Henderson, and A. Levin. 2000. Optimal monetary policy with staggered wage and price contracts. *Journal of Monetary Economics* 46: 281–313.

Fisher, I. 1920. *Stabilizing the dollar*. Norwood, MA: Norwood Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1): 1–17.

Gali, J. 1999. Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review* 89: 249–271.

Humphrey, T. 2004. Classical deflation theory. *Federal Bank of Richmond Economic Quarterly* 90: 11–32.

Keynes, J. 1936. *The general theory of interest, employment, and money*. London: Macmillan.

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

Phelps, E. 1968. Money–wage dynamics and labor market equilibrium. *Journal of Political Economy* 76: 678–711.

Sargent, T., and N. Wallace. 1975. Rational expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83: 169–183.

Taylor, J. 1999. Staggered wage and price setting in macroeconomics. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford. Amsterdam: North-Holland.

Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37: 345–370.

M

Monetary Business Cycles (Imperfect Information)

Christian Hellwig

Abstract

Business cycle theories based on incomplete information start from the premise that key economic decisions on pricing, investment or production are often made on the basis of incomplete knowledge of constantly changing aggregate economic conditions. As a result, decisions tend to respond slowly to changes in economic fundamentals, and small or temporary economic shocks may have large and long-lasting effects on macroeconomic aggregates. This article provides an introductory overview of incomplete information-based theories of business cycles, from their origins to the most recent theoretical developments.

Keywords

Common information; Forecasting the forecasts of others; Heterogeneity in beliefs; Higher-order expectations; Higher-order uncertainty; Imperfect common knowledge; Imperfect information; Information aggregation; Informational frictions; Law of iterated expectations; Local markets; Menu cost; Monetary business cycle models; Monetary shocks; Monopolistic competition; New Keynesian economics; Pricing complementarities; Rational expectations; Rational inattention; Real rigidities; Sticky prices

JEL Classifications

D4; D10

Business cycle theories based on incomplete information start from the premise that key economic decisions on pricing, investment or production are often made on the basis of incomplete knowledge of constantly changing aggregate economic conditions. As a result, decisions tend to respond slowly to changes in economic fundamentals, and small or temporary economic shocks may have large and long- lasting effects on macroeconomic aggregates.

Incomplete information theories have been popular in particular for explaining sluggish price or wage adjustment in response to monetary shocks. At the heart of this theory lies the assumption that firms or households only pay attention to a relatively small number of indicators regarding conditions in markets relevant to their own activities, but they may not acquire information more broadly about aggregate economic activity. With imprecise information about these aggregate conditions, it takes the firms some time to sort out temporary from permanent changes, or nominal from real disturbances. Prices then respond with a delay to changes in nominal spending, and monetary shocks may have significant effects on real economic activity in the intervening periods – despite the fact that firms have the opportunity to constantly readjust their decisions.

This basic idea was proposed first by Phelps (1970) and formalized by Lucas (1972). In Lucas (1972), economic agents produce in localized markets, in which they observe the market-clearing price at which they can sell their output. This price is affected both by aggregate spending shocks and by market-specific supply shocks. Under perfect information, quantities adjust in response to local supply shocks, but not prices, and prices respond to aggregate spending shocks, but not quantities. With imperfect information, agents are unable to filter out the magnitudes of the aggregate and market-specific shocks from the observed prices in the short run. Output then responds positively to price changes and spending shocks in the short run, but not in the long run, once agents have been able to sort out the spending shocks from the market-specific supply shocks.

Lucas (1972) formulated this idea in a rational expectations market equilibrium model, in which agents' expectations are fully Bayesian, and the resulting output responses are optimal. His model also includes stark assumptions about the nature of local versus aggregate market interactions, as well as the nature of shocks (monetary versus real, demand versus supply, aggregate versus market-specific) and the information to which firms have access.

Importantly, the model lacks a natural internal amplification mechanism: the extent of incomplete nominal adjustment depends almost entirely on the degree of informational incompleteness. Subsequent work has tried to address these issues, for example by introducing richer information structures. Townsend (1983) considers an investment model in which firms get to observe how much some of the other firms invest. Therefore, they need to form forecasts about each others' beliefs – forecasting the forecasts of others. This leads to a complicated infinite regress problem, whereby a firm's current investment level depends on its observation of other firms' past investment, which in turn depended on observations about past investment... Townsend showed that this type of problem does not admit a simple finite-dimensional recursive structure. As a result, firms must draw inference about all past realizations of shocks simultaneously, leading to an infinite-dimensional filtering and fixed point problem, with no easily characterized solution.

These and other important technical and computational hurdles effectively imposed limitations on the complexity and economic realism of the early incomplete information models. Moreover, the model is open to the criticism that if incomplete information is a major source of business cycle fluctuations, then there seems to be an important societal benefit to making the relevant information publicly available to everyone. In part because of these difficulties, economists have, from the mid-1980s, turned their attention to New Keynesian sticky price theories that emphasize the role of adjustment and coordination frictions in price-setting. (Among others, see Calvo 1983; Blanchard and Kiyotaki 1987.)

Recently, the incomplete information theories have made a comeback, which can be traced to two factors. First, technological progress has made models such as Townsend (1983) computationally tractable. Second, new game-theoretic results regarding equilibrium analysis with a lack of common knowledge and heterogeneity in beliefs, as well as insights borrowed from the sticky price literature regarding the role of real rigidities and pricing complementarities (Ball and Romer 1990) have enabled us to paint a much richer picture of the adjustment dynamics resulting from incomplete information models. The empirical performance of these new incomplete information models, however, still remains to be seen.

In the remainder of this article I provide a unified exposition of the main ideas behind the incomplete information theories, from the original contributions to the more recent renewal. I also attempt to chart out some of the challenges that lie ahead. This is a lively and active area of research, with many open questions and few definite answers.

A Canonical Framework

Consider the following model, which is based on the New Keynesian models of monopolistic competition. There is a large number of firms, indexed by $i \in [0, 1]$. In each period, each firm sets its (log-)price $p_t(i)$ equal to its expectation of a target price $p_t^*, p_t(i) = E\left(p_t^* \mid \mathcal{I}_t^i\right)$, where $\mathcal{I}_t^i$ denotes the information set of firm i at date t, that is all signals on which it can condition its pricing decision. p_t^* is characterized as

$$p_t^* = ky_t + p_t, \tag{1}$$

where $p_t = \int p_t(i)di$ denotes the average of the firms' pricing decisions, y_t denotes the aggregate real output in period t, relative to its trend level that would prevail with complete information, and $k > 0$ measures the response of optimal pricing decisions to real output. A firm's ideal relative price $p_t^* - p_t$ is determined by real output deviations from trend.

We augment this pricing rule by a quantity equation, $y_t + p_{t.} = m_t$, where m_t denotes nominal spending. Combining the two, we find

$$p_t^* = km_t + (1 - k)p_t. \tag{2}$$

Nominal spending m_t is driven by exogenous shocks; for simplicity, assume that $m_t = m_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}$ is i.i.d. white noise.

Each firm's target price is therefore a linear combination of the exogenous shocks and the prices set by the other firms. If $k \in (0, 1)$, prices are complementary, that is, an increase in the average price level implies that each firm has an incentive to raise its own price. The parameter value of k depends on the substitution elasticity between the firms' products, the firms' returns to scale parameter in the technology, and the Frisch elasticity of labour supply.

To complete the model description, we need to specify each firm's information set $\mathcal{I}_t^i$ – this is where different incomplete information theories vary. An equilibrium of this model requires that prices satisfy the optimality condition $p_t(i) = E\left(p_t^* \mid \mathcal{I}_t^i\right)$, taking into account that p_t^* itself depends on the aggregate price level.

Common Information

Suppose first that all firms have identical information sets, $\mathcal{I}_t^i = \mathcal{I}_t$. Then, they will set identical prices, equal to $p_t(i) = p_t = E(m_t \mid \mathcal{I}_t)$. This reflects the implications of the original Lucas

M

model that prices adjust to the common expectation of the underlying shocks. When information is incomplete, firms will only learn gradually about m_t, prices adjust slowly, and monetary surprises have real effects: y_t is determined directly by the discrepancy between the realized and the expected value of m_t. However, if the available information on which these expectations are based is sufficiently precise, then $E(m_t|\mathcal{I}_t)$ cannot be far from the true value of m_t. As discussed above, the real effects of monetary shocks are bounded by the degree of informational incompleteness – as firms have better information, their prices track m_t more closely, and monetary shocks have smaller real effects.

Heterogeneous Beliefs, but Independent Strategies

A similar conclusion emerges when firms have different information sets, but their target prices do not respond to the other firms' decisions ($k = 1$). Each firm's price is set equal to its expectation of the spending shock $p_t(i) = E(m_t|\mathcal{I}_t^i)$, and the average price adjusts according to the average expectation $p_t = \overline{E}(m_t) = \int E(m_t|\mathcal{I}_t^i)\,di$ of the spending shock. Once again, if firms are sufficiently well informed, their pricing decisions will on average not be far from the nominal spending shock, which implies little delay in price adjustment and only small real output effects.

Heterogeneous Beliefs and Complementary Strategies

Suppose now that instead $k \in (0, 1)$, so that there are complementarities in pricing decisions. Averaging the pricing equation, and substituting forward, firm i's equilibrium price is given by

$$p_t(i) = k\sum_{s=0}^{\infty}(1-k)^s E\left[\overline{E}^{(s)}(m_t)|\mathcal{I}_t^i\right] \quad (3)$$

where $\overline{E}^{(s)}(m_t)$ denotes the s-order average expectation of m_t, or the average expectation of the average expectation of... (repeat s times) ... of m_t. A firm's optimal price is therefore given as a geometrically weighted average of higher-order expectations – a firm needs to forecast not only the realized shock but also the other firms' expectations of the shock, the other firms' expectations of the other firms' expectations of the shock, and so on.

If the firms all had identical information, the law of iterated expectations would simply collapse the right-hand side above into the common first-order expectation of m_t. The model thus derives its interest from the fact that with heterogeneous information, higher-order expectations respond differently to new information than first-order expectations about m_t.

The following example illustrates this point and serves also to derive the main results of this model. Suppose that all firms observe m_{t-1} exactly, but only a fraction λ (the *informed*) gets to observe m_t. Then, $\overline{E}(m_t) = \lambda m_t + (1-\lambda)m_{t-1}$, but the second order average expectation is

$$\begin{aligned} E^{(2)}(m_t) &= \lambda[\lambda m_t + (1-\lambda)m_{t-1}] + (1-\lambda)m_{t-1} \\ &= \lambda^2 mt + \left(1-\lambda^2\right)m_{t-1}. \end{aligned}$$

By iteration, the s-order average expectation of m_t is $\overline{E}^{(s)}(m_t) = \lambda^s m_t + (1-\lambda^s)m_{t-1}$. The average price is

$$\begin{aligned} p_t &= k\sum_{s=0}^{\infty}(1-k)^s\overline{E}^{(s+1)}(m_t) \\ &= m_{t-1} + \frac{k\lambda}{1-(1-k)\lambda}(m_t - m_{t-1}). \end{aligned} \quad (4)$$

Two important conclusions emerge. First, note that $\frac{k\lambda}{1-(1-k)\lambda} < \lambda$. The informed firms whose prices may react to m_t take into account that the uninformed firms won't respond, which in turn reduces their incentives to adjust prices. Therefore, while incomplete information serves as the initial source of sluggish price adjustment, the complementarity and the heterogeneity in beliefs dampen the response of prices far beyond what the initial degree of informational incompleteness would suggest. To illustrate the strength of this amplification effect, consider the following

numerical example: suppose that $k = 0.15$ (as in standard parametrizations of New Keynesian sticky price models), and that half the firms are informed. Then, the contemporaneous response of average prices is $\frac{k\lambda}{1-(1-k)\lambda} \approx 0.13$, that is a 1 per cent increase in nominal spending leads to only a 0.13 per cent increase in prices, and a 0.87 per cent increase in real output – despite the fact that half of the firms actually observe the increase in nominal spending and are hence able to respond to it!

Second, this amplification can be large, even if the degree of informational incompleteness is small. If λ is close to 1, almost all firms exactly observe the current realization m_t. Nevertheless, if k is close to 0, that is if there is a strong pricing complementarity, they still won't respond to the monetary shock. The presence of only a few uninformed firms is therefore enough to radically overturn the conclusions of the complete information model.

These two observations apply quite generally, once firms have heterogeneous beliefs. They form the central insight of the new incomplete information theories. In Mankiw and Reis (2002), heterogeneous beliefs result because, in any given period, only a fraction of firms observe new information. This generalizes the above example to allow for richer adjustment dynamics. In Woodford (2002), all firms observe a conditionally independent idiosyncratic signal x_t^i of the current realization of m_t in each period. The resulting inference problem is more complicated but can be solved numerically. Again, the response of prices to monetary shocks is significantly dampened by the fact that firms do not share in common information, yet their pricing decisions are complementary.

The Role of Public Information

Hellwig (2002) provides a simplified version of Woodford (2002), providing closed-form solutions to a general class of information structures. This simplified model also accommodates the presence of additional public sources of information such as central bank announcements. Besides dampening the response to idiosyncratic private signals, the complementarity in prices generates overreaction to public news. Public announcements thus speed up price adjustment and reduce the real effects of monetary shocks, but the noise in public news creates an additional source of volatility, which in some cases may increase rather than decrease real output fluctuations. (Similar results are derived by Amato and Shin (2003) for Woodford's model, and by Ui (2003) in the original Lucas island model.)

Looking Ahead

These new contributions have provided promising insights into the amplification and propagation mechanisms of incomplete information models. But they also abstract from important modelling issues that need to be addressed before a comprehensive quantitative evaluation becomes possible.

So far, much of the analysis is based on a stylized price-setting model that captures the essence of pricing complementarities as described above, without deriving them within a fully specified dynamic general equilibrium model. This short-cut is not without problems. First, the lack of a proper context of markets makes it difficult to interpret these propagation results. Presumably in a market firms obtain some information about price and quantity variables – so far, this is not formally modelled.

Second, the assumption that firms are heterogeneously informed implies that other frictions must be present – in particular, the extent to which information about fundamental shocks can be inferred from publicly observable prices must be limited, implying that the asset market must be incomplete. But then, one faces the problem of isolating the effects of informational heterogeneity from the effects of other market imperfections. In Lorenzoni (2006) for instance, a precautionary savings motive generates a multiplier effect in household spending, which is further amplified by the presence of heterogeneous information.

Third, there is an issue of interpretation. At this point, there exist several different interpretations

regarding the source of the differences in beliefs across firms, and they may lead to radically different model conclusions. In Mankiw and Reis (2002), firms update their information only infrequently, and in the intervening periods set prices on the basis of outdated information; Reis (2006) further develops this idea on the basis of menu costs in updating decisions. Woodford (2002) instead bases his model on the notion of 'rational inattention', developed by Sims (2003, 2006a). Sims argues that decision makers only have a finite capacity to process new information, which constrains the quality of the signals they observe in any given period. Heterogeneity in beliefs then arises naturally through the idiosyncratic noise in each individual's information processing channel (see Sims 2006b, for further discussion of the resulting conceptual and modelling issues). A third interpretation suggests that individuals are Bayesian, but access to information is limited – for example, firms observe the demand for their own products, but not the demand for competitors' products. If each firm is subject to idiosyncratic, as well as common shocks, then an information structure much like the above with idiosyncratic private signals emerges. On the other hand, firms also observe market prices, which generates a source of common information.

Finally, all these models treat the information structure as an exogenous primitive. In reality, firms and households have access to overwhelming amounts of information, and information processing becomes a matter of choice, given the existing constraints and trade-offs. By and large, the effects of information costs and choices and the strategic interaction that results from these choices remains unexplored. Preliminary developments in this direction include Mackowiak and Wiederholt (2005) and Hellwig and Veldkamp (2005). In Mackowiak and Wiederholt, firms need to allocate a fixed information processing capacity between firm-specific and aggregate variables. Hellwig and Veldkamp explore how the pricing complementarities that are relevant for business cycle implications also shape incentives for information acquisition.

In summary, the most important issue that remains to be resolved is the grounding of new incomplete information theories within a fully specified model of goods and asset markets, with special emphasis on the origins of the informational frictions. Beyond that, the new incomplete information theories raise many intriguing questions, which merit further attention, or have already been addressed to some extent: for example, Ball et al. (2005) reconsider the role of monetary policy, and Morris and Shin (2002), Hellwig (2005) and Angeletos and Pavan (2004, 2007) discuss the welfare effects of information disclosures. Finally, the combination of new evidence on the cross-sectional and business cycle properties of expectations (Mankiw et al. 2004) and new micro-level data on price adjustments (Bils and Klenow 2004) promises to provide an interesting avenue for evaluating the empirical performance of the model's cross-sectional and business cycle implications.

See Also

- ▶ Information Aggregation and Prices
- ▶ Lucas, Robert (Born 1937)
- ▶ Monetary Business Cycle Models (Sticky Prices and Wages)
- ▶ Monetary Transmission Mechanism
- ▶ Phelps, Edmund (Born 1933)

Bibliography

Amato, J., and H.S. Shin. 2003. *Public and private information in monetary policy models*. Working Paper No. 138, Bank of International Settlements.

Angeletos, G.-M., and A. Pavan. 2004. Transparency of information and coordination in economies with investment complementarities. *American Economic Review* 94 : 91–98.

Angeletos, G.-M., and A. Pavan. 2007. Efficient use of information and social value of information. *Econometrica* 75 : 1103–1142.

Ball, L., and D. Romer. 1990. Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57 : 183–203.

Ball, L., G. Mankiw, and R. Reis. 2005. Monetary policy for inattentive economies. *Journal of Monetary Economics* 52 : 703–725.

Bils, M., and P. Klenow. 2004. Some evidence on the importance of sticky prices. *Journal of Political Economy* 112 : 947–985.
Blanchard, O., and N. Kiyotaki. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77 : 647–666.
Calvo, G. 1983. Staggered prices in a utility maximizing framework. *Journal of Monetary Economics* 12 : 383–398.
Hellwig, C. 2002. *Public announcements, adjustment delays and the business cycle*. Discussion paper, University of California, Los Angeles.
Hellwig, C. 2005. *Heterogeneous information and the welfare effects of public information disclosures*. Discussion paper, University of California, Los Angeles.
Hellwig, C., and L. Veldkamp. 2005. *Knowing what others know: Coordination motives in information acquisition*. Discussion paper, University of California, Los Angeles and New York University.
Lorenzoni, G. 2006. *A theory of demand shocks*. Discussion paper, Massachusetts Institute of Technology.
Lucas, R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4 : 103–124.
Mackowiak, B., and M. Wiederholt. 2005. *Optimal sticky prices under rational inattention*. Discussion paper, Humboldt University Berlin.
Mankiw, G., and R. Reis. 2002. Sticky information versus sticky prices: A proposal to replace the new Kaynesian Phillips curve. *Quarterly Journal of Economics* 117 : 1295–1328.
Mankiw, G., R. Reis, and J. Wolfers. 2004. Disagreement about inflation expectations. In *NBER macroeconomics annual 2003*. Cambridge, MA: MIT Press.
Morris, S., and H.S. Shin. 2002. The social value of public information. *American Economic Review* 92 : 1521–1534.
Phelps, E. 1970. Introduction: The new microeconomics in employment and inflation theory. In *Microeconomic foundations of employment and inflation theory*. New York: Norton.
Reis, R. 2006. Inattentive producers. *Review of Economic Studies* 73 : 1–29.
Sims, C. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50 : 665–690.
Sims, C. 2006a. Rational inattention: Beyond the linear-quadratic case. *American Economic Review* 96 : 158–163.
Sims, C. 2006b. *Rational inattention: A research agenda*. Discussion paper, Princeton University.
Townsend, R. 1983. Forecasting the forecasts of others. *Journal of Political Economy* 91 : 546–588.
Ui, T. 2003. *A note on the Lucas model: Iterated expectations and the non-neutrality of money*. Discussion paper, Yokohama National University.
Woodford, M. 2002. Imperfect common knowledge and the effects of monetary policy. In *Knowledge, information and expectations in modern macroeconomics*, ed. P. Aghion et al. Princeton: Princeton University Press.

Monetary Cranks

David Clark

Keywords
Douglas, C. H.; Gold standard; Monetary cranks; Money supply; Social credit; Underconsumptionism

JEL Classifications
O

The history of ideas tends to concentrate on the successful ideas – ideas which appear to have been precursors of the orthodoxy of the day. As a result, ideas which had large followings but which are later considered 'cranky' tend to be ignored. This is especially true of the ideas of those who we can loosely call the monetary cranks.

These persons have placed money at the centre of their economic analysis, have usually placed major blame for society's evils on alleged financial conspiracies and bankers' ramps – on the 'Money Power' – and have advocated a variety of monetary experiments. Over the past century particularly, such concerns can be found in all Western countries, on both the Left and the Right of politics. This article can only provide the broadest of overviews of the voluminous literature in this field.

Opposition to financial oligarchies has a long history. The Medicis of 15th-century Florence aroused suspicion and hostility. In *Lombard Street* (1873), Walter Bagehot described the streets around the Bank of England in London as 'by far the greatest combination of power and economic oligarchy that the world has ever seen'. But it was the fiery late-19th-century American populist, William Jennings Bryan, who popularized the term 'Money Power' (cited in Douglas 1924, Preface):

> The Money Power preys upon the nation in times of peace and conspires against it in times of adversity. It is more despotic than monarchy, more insolent

> than autocracy, more bureaucratic than bureaucracy. It denounces, as public enemies, all who question its methods, or throw light upon its crimes. It can only be overthrown by the awakened conscience of the nation.

Monetary parables have a long history, ranging from David Hume's 1752 hope that 'by miracle, every man in Great Britain should have five pounds slipped into his pocket in one night', through to Milton Friedman's 1969 postulated helicopter miracle, whereby dollars would be dropped from the heavens. (These are discussed in Clayton et al. 1971, p. 6.) Over the past three centuries, however, actual monetary experiments have taken two main forms: attempts to overcome economic fluctuations by means of adjusting note issue; and attempts to achieve a more stable price level through the formulation and adoption of a new or different monetary standard.

Such experiments were first undertaken in the North American colonies. The first paper money issued by any government in Europe or the Americas was printed by Massachusetts to pay the wages of its soldiers engaged in conflict with the French in Canada at the end of the 17th century. Other New England colonies followed suit and a competitive depreciation of the individual currencies followed. The French Canadians even used playing cards as a form of money.

In 1721, a Mr Wise of Chebacco, Massachusetts, concerned at the depreciation of the notes admonished his fellow colonists (cited in Lester 1939):

> Gentleman! You must do by your Bills, as all Wise Men do by their Wives; Make the Best of them... Wise Men Love their Wives; and what ill- conveniences they find in them they bury; and what Vertues they are inrich't with they Admire and Magnifie. And thus you must do by your Bills for there is not doing without them; if you Divorce or Disseize yourselves of them you are undone.

Hence the American colonies developed the practice of adjusting note issue to stimulate business or countervail a recession. They believed that there is a very close relationship between money, prices and business conditions and that the appropriate note issue would greatly stimulate business. Their efforts were made easier by the fact that there was no bank-issued money.

In England, after the Napoleonic Wars, the first great debate about monetary reform occurred, with persons such as Joseph Lowe, John Rooke and Poulett Scrope, proposing a 'managed currency', the volume of which was to be controlled according to changing prices in such a way as to keep the price level steady. Similarly, Henry Thornton's *Paper Credit* (1802) argued that contraction or expansion of the money supply had real effects on the level of economic activity. In the 1840s, Thomas Attwood claimed that if Britain's coinage 'were accommodated to man and man to our coinage then world would be capable of multiplying its production to an unlimited extent'. However, David Ricardo's and John Stuart Mill's failure to appreciate that credit expansion might stimulate the level of economic activity, rather than just increase prices, dominated economic thinking for the rest of the 19th century (see Viner 1937).

This opened the door for the monetary cranks, who argued that money did matter. Their main inspiration came from the underconsumptionist tradition. A number of authorities have emphasized that underconsumptionist literature is difficult to categorize (for example, Schumpeter 1954, p. 740; Haberler 1937, chapter 5; Bleaney 1976, chapter 1). Still, the argument that there is a permanent deficiency of purchasing power produced all kinds of suggestions as to how such a deficiency could be remedied.

In the interwar period, underconsumptionist ideas fell on particularly receptive ears. Many persons, particularly those concerned with high unemployment, were prepared to believe that the schemes of the monetary cranks would increase demand and hence create jobs. The quantity of pamphlet literature on monetary reform over this era is thus enormous. A common argument was that because the First World War was financed by printing money, the same method could be used to eliminate unemployment. Opposition to the gold standard usually accompanied this argument.

Academic discussion of monetary matters was disparate and disputatious (see, for example the famous debate between F.A. von Hayek and P. Sraffa in the *Economic Journal,* March–June 1932) and this was seized upon by the monetary

reformers, who sought to penetrate what they claimed were the obfuscations of the academics. They also pointed to the fact that discussion of money and banking tended to be confined to tendentious tomes written for bank employees, while economic theory textbooks devoted little space to arguments against Say's Law.

Major C.H. Douglas was probably the best-known reformer in English-speaking countries in this era (see Douglas 1924) but there were many, many others who wrote on monetary reform. These included: A.H. Abbati, who attracted the interest of John Maynard Keynes and D.H. Robertson; Sir Normal Angel, whose set of cards *The Money Game* was widely used in high schools in Britain and the US; W.T. Foster and W. Catchings, who were probably the best known US reformers; and Frederick Soddy of Oxford University, who, after being awarded the Nobel Prize for chemistry, set out to solve the money problem inspired by John Ruskin's *Unto this Last* (1862) and an Australian invention. Soddy argued that the gold standard could be replaced with a machine based on the automatic totalizator at Sydney's Randwick Racecourse (Soddy 1931). Cole (1933) discusses some of this literature.

Strangely, Schumpeter (1954) contains no reference to Douglas but he does mention (pp. 1090–91) G.F. Knapp's *The State Theory of Money* (1924), which promoted similar ideas and had considerable impact in interwar Germany. For example, in the dying days of the Weimar Republic, at the suggestion of H.J. Rustow and W. Lavtenbach of the Ministry of Economics, interest-bearing tax certificates were issued in lieu of treasury bills and exchequer bonds. Employers were given these certificates if they employed additional employees and reduced the wages of existing employees (see Rustow 1978).

With the Keynesian revolution and the increased emphasis given to monetary theory by academic economists in recent decades, the monetary cranks have largely disappeared from public debate, although underconsumptionist ideas will probably have supporters while ever there is unemployment.

Any explanation of the appeal of these ideas over generations would have to invoke sociology and psychology. Such ideas found strong support because they enabled persons to impress their peers with their apparent understanding of economics, even though they had no formal training in the discipline. They offered the false hope that there were simple solutions to the complexities of modern economic life. They also transcended party political allegiances – similar passages about 'credit slavery' and 'Shylocks' can be found in Hitler's *Mein Kampf* and leftwing pamphlets of the same era. A very wide range of individuals can be opposed to private banks and the 'Money Power' without their opposition leading to more sophisticated political analysis. In fact, as the history of populism shows, 'Funny Money' beliefs provided a kind of ideological release valve.

The history of ideas contains numerous examples of the power of the phrasemonger. The simpler the panacea, the greater the chance the agitator will have of attracting a following. As the Chartist agitator Ernest Jones once advised (cited in Martin and Rubinstein 1979, p. 43): 'We say to the great minds of the day, come among the people, write for the people and your fame will live forever.'

Bibliography

Angell, N. 1936. The money mystery: An explanation for beginners. London: Dent (*The Money Game*, a set of cards for teaching purposes, was sold in conjunction with this book).

Bleaney, M. 1976. *Underconsumption theories: A history and critical analysis*. London: Lawrence & Wishart.

Clayton, G., J.C. Gilbert, and R. Sedgwick, eds. 1971. *Monetary theory and policy in the 1970s*. London: Oxford University Press.

Cole, G.D.H. 1933. *What everybody wants to know about money: A planned outline of monetary problems*. London: Victor Gollancz.

Douglas, C.H. 1924. *Social credit*. London: Eyre & Spottiswoode.

Durbin, E.F.M. 1934. *Purchasing power and trade depression*. London: Chapman & Hall.

Haberler, G. 1937. *Prosperity and depression*. Cambridge: Harvard University Press.

Lester, R.A. 1939. *Monetary experiments: Early American and recent Scandinavian*. Princeton: Princeton University Press.

Martin, D., and D. Rubinstein, eds. 1979. *Ideology and the labour movement*. London: Croom Helm.

Rustow, H.J. 1978. The economic crisis of the Weimar Republic and how it was overcome. *Cambridge Journal of Economics* 2: 409–421.
Schumpeter, J.A. 1954. *A history of economic analysis*. London: George Allen & Unwin.
Soddy, F. 1931. *Money versus man*. London: Elkin Mathews & Marrot.
Viner, J. 1937. *Studies in the theory of international trade*. London: George Allen & Unwin.

Monetary Disequilibrium and Market Clearing

Herschel I. Grossman

Conventional wisdom interprets the empirical relation between monetary aggregates and measures of real aggregate economic activity primarily as reflecting the effect of monetary policy on real activity. A host of historical episodes apparently accord with this interpretation. It is, for example, hard to deny that disinflationary monetary policy contributed to the 1982 recession in the United States.

Some theorists, such as King and Plosser (1984), have questioned this interpretation and have developed real business cycle models that attempt to explain the observed correlations of money and real activity as solely a result of the common influences of other factors, such as disturbances to tastes, technology, and resources or disturbances to monetary velocity. These theorists, however, have not been able to identify an alternative set of impulses that does not contain disturbances to monetary aggregates and that does have appropriate structural characteristics, sufficient magnitude, and requisite regularity to be responsible for the bulk of observed fluctuations in real activity. This inability to identify alternative causal factors reinforces the standard reading of history that monetary policy influences real activity. (See McCallum (1986) for a thorough critique of real business cycle models.)

Given the conventional interpretation of the observed relation between money and real activity, a satisfactory theoretical and empirical analysis of macroeconomic fluctuations must account for an effect of monetary policy on real activity as well as for an effect of monetary policy on inflation. This account must be consistent with the following general features of the data: (1) current realizations of monetary aggregates are correlated with subsequent realizations of both real activity and inflation; (2) the correlations of money with real activity are strong in the short run but weaken in the long run whereas the correlations of money with inflation are weak in the short run but become stronger in the long run; and (3) the correlations with real activity are stronger for unanticipated realizations of monetary aggregates whereas the correlations with inflation are stronger for anticipated realizations of monetary aggregates. The main attraction of monetary-disequilibrium theory, which is the useful name that Leland Yeager (1986) uses for what is often called the Keynesian or non-market-clearing approach, is that it provides an explanation for the effects of monetary policy on real activity and inflation that in its modern versions, which incorporate the natural-rate hypothesis and the rational-expectations hypothesis, seems to be broadly consistent with these general features of the data.

An explanation for the effect of monetary policy on real activity also must satisfy criteria of logical consistency. Most importantly, aggregate economic activity is merely a statistical summary of a multitude of individual productive decisions, which are the same individual decisions that determine resource allocation and income distribution. Accordingly, the assumptions about economic behaviour used to account for the relation between money and real activity should be consistent with the assumptions used to explain resource allocation and income distribution. Moreover, we cannot avoid this consistency requirement by asserting that macroeconomic fluctuations are a short-run phenomenon, whereas questions about resource allocation and income distribution involve the long run. In fact, economists routinely apply standard microeconomic analysis to the

short run – that is, to a time horizon shorter than the typical business cycle.

The distinguishing feature of conventional economic analysis of resource allocation and income distribution is the assumption that producers in free markets exhaust perceived opportunities for mutually advantageous exchange. Standard microeconomic analysis takes this assumption to be a corollary of the basic economic postulate of maximization. The most unattractive aspect of monetary-disequilibrium theory is that, as yet, its proponents (who include most macroeconomists) have been unable to reconcile it with the postulate of maximization and the corollary that perceived gains from trade are exhausted.

A frequent claim is that the existence of coordination problems reconciles monetary disequilibrium with the postulate of maximization. Various authors argue that, even with producers behaving as rational maximizers, perception and coordination of the wage and price adjustments necessary to clear markets in the face of unanticipated monetary disturbances takes time. For example, Yeager (1986) points out that 'one cannot consistently both suppose that the price system is a communication mechanism – a device for mobilizing and coordinating knowledge dispersed in millions of separate minds – and suppose that people *already* have the knowledge that the system is working to convey.' This observation is correct, but it seems irrelevant for the analysis of monetary disequilibrium because the values of monetary aggregates are public information. In contrast to truly private information, the monetary aggregates are not information that the price system has to convey.

A further frequent claim is that even with complete information, strategic considerations would cause individual rationality to diverge from the collective rationality implicit in monetary equilibrium. In his Presidential Address to the American Economic Association, Charles Schultze (1985) invokes the analogy of the prisoner's dilemma to argue that the unwillingness of any producer 'to go first' would inhibit wage and price adjustments. This analysis is confusing because it seems to imply too much – namely, that wages and prices are rigid rather than merely stickly. In any event, the usefulness of the prisoner's dilemma analogy for understanding market behaviour seems limited because the prisoner's dilemma relates to a hypothetical game played by a small number of agents who cannot communicate with each other during the game.

For a monopolist or collusive oligopoly, individual and collective optimality of wage and price adjustments obviously coincide. In a market of many imperfectly competitive producers, however, optimal individual wage and price responses to some disturbances can differ from optimal collective responses. But observed changes in monetary aggregates are not such a disturbance. Unless price adjustments are prohibitively costly, optimal individual price setting behaviour requires responding to an observed disturbance of monetary aggregates even if the individual thinks that other individuals are ignoring the disturbance. The 'initial' response, of course, might not be an equiproportionate price adjustment but, even without rational expectations, subsequent responses culminate in an equi-proportionate adjustment. Moreover, if we assume either that expectations are rational or that price-adjustment costs are small, the theory suggests that the full adjustment is essentially instantaneous.

Schultze and Yeager also refer to models of efficient long-term contracts and implicit buyer-seller understandings. This reference is puzzling, because, although these models suggest that real or relative wages and prices would be less flexible than models of spot markets imply, models of efficient contracts also suggest, if anything, that rational wage setters would fully index nominal wages and prices to observed monetary disturbances. Schultze recognizes this point, but claims that the complexity of the relation between monetary aggregates and marketclearing nominal wages precludes indexation. It is not clear, however, why this problem results in zero indexation. Even if producers cannot easily determine the optimal degree of indexation, they surely know that some positive indexation would be better than zero indexation. Similarly, currently popular models of efficiency wages, whatever their ability

to explain the equilibrium structure of real wages and employment, also have no apparent relevance for the problem of rationalizing stickiness of nominal wages and resulting monetary disequilibrium.

In the early 1970s, theorists like Robert Lucas (1972, 1973) and Robert Barro (1976) responded to the problem of reconciling monetary disequilibrium with the postulate of maximization by utilizing advances in the theory of expectations and general economic equilibrium under incomplete information to formulate 'equilibrium' models of macroeconomic fluctuations. These equilibrium models assume that all perceived gains from trade are realized and that expectations are rational, and they rely on assumed lack of information about monetary aggregates in order to generate an effect of monetary aggregates on real activity. In recent years, interest in these equilibrium models has waned largely because more extensive theoretical and econometric analysis has shown these models to be unable to account for the observed relation between monetary aggregates and real activity.

The empirical problem with equilibrium models, it should be stressed, does not involve direct evidence that perceived gains from trade are actually not realized. In fact, contractual versions of equilibrium models – see, for example, Azariadis (1978) and Grossman (1981) – readily account for prominent observed features of macroeconomic fluctuations that would seem inconsistent with market clearing if market clearing were narrowly interpreted in a framework of spot markets. These observed features include lack of correlation between aggregate employment and real wage rates and the use of layoffs to effect employment separations.

The empirical rejection of equilibrium models is based on rejection of an essential testable implication of the combined assumptions that all perceived gains are realized and that expectations are rational. This implication is that disturbances to monetary aggregates affect real aggregates only to the extent that currently available information does not permit agents to infer current monetary aggregates accurately. The testable form of this implication, derived by Boschen and Grossman (1982) following the lead of King (1981), is that the current innovation in real activity is uncorrelated with contemporaneous measures of current and past changes in monetary aggregates. Not surprisingly, econometric analysis of data for the United States reported by Boschen and Grossman not only unambiguously rejects this hypothesis, but also finds no correlation between the innovation in real activity and revisions in preliminary estimates of monetary aggregates, these revisions being measures of the unperceived part of monetary policy.

The early equilibrium models of Lucas and Barro obscured the problem of reconciling equilibrium assumptions with the observed relation between monetary aggregates and real activity because they abstracted from the existence of contemporaneously available monetary data. Barro himself was among the first to recognize the consequences of relaxing this abstraction. An empirical study by Barro and Hercovitz (1980) anticipated the subsequent and more formal theoretical and econometric analysis of King and Boschen and Grossman. In an early reassessment of equilibrium theories, Barro wrote,

> A significant weakness of the [equilibrium] approach is the dependence of some major conclusions on incomplete contemporaneous knowledge of monetary aggregates, which would presumably be observed cheaply and rapidly if such information were important. The role of incomplete current information on money in equilibrium business cycle theory parallels the use of adjustment costs to explain sticky wages and prices with an associated inefficient determination of quantities in Keynesian models. The underpinning of the two types of macroeconomic models are both vulnerable on a priori grounds . . . (Barro 1981a, ch. 2, p. 74)

On the same page, however, Barro is quick to emphasize that doubts about the explanatory value for business cycles of currently available equilibrium theories do not constitute support for Keynesian disequilibrium analysis. The disequilibrium theories are essentially incomplete models that raise even larger questions about the consistency of model structure with underlying rational behaviour. It remains a fair observation that existing macroeconomic theories – including

new and old approaches – provide only limited knowledge about the nature of business cycles.

Lucas also has recognized the consequences for the implications of equilibrium models of taking contemporaneous monetary information into account. In a recent lecture Lucas (1985) acknowledges that 'insofar as the monetary information necessary to permit agents to correct for what are, or ought to be, units changes is public . . . then one would expect this information to be used, independent of the form of interaction among agents.' Nevertheless, Lucas still seems willing to defend abstracting from contemporaneous monetary data as an 'as-if' assumption, although apparently he can only vaguely conjecture why rational agents would ignore information that is important and freely available. In the same lecture, he offers only the thought that 'it seems to me most unlikely that it would be in the private interest of individual agents to specialize their individual information systems so as to be well-equipped to adapt for units changes of monetary origin.'

As an alternative to the formulations of equilibrium models, other theorists have reacted to the difficulty of reconciling monetary disequilibrium with the postulate of maximization by appealing, either implicitly or explicitly, to concepts of near rationality. The seminal work of Stanley Fischer (1977), incorporating rational expectations into a non-market-clearing framework, is an important example of this approach. In Fischer's model, although nominal wages are sticky, these predetermined nominal wages are equal to rational expectations of market-clearing wages.

Econometric testing of these nearly rational, monetary-disequilibrium models with rational expectations encounters the difficult problem of realistically dating the formation of the expectations relevant for the determination of current nominal wages and current real activity. As explained in Grossman (1983), Barro's empirical results on the relation between real activity and unanticipated monetary disturbances, summarized in Barro (1981b), provide qualified support for Fischer's model. In another study, Grossman and Haraf (1985), by taking advantage of the fact that wage setting in Japan is both decentralized and synchronized, were able to examine empirically some detailed implications of Fischer's model and to show that the model, if suitably elaborated, seems to fit the Japanese data.

More recent theoretical work by Akerlof and Yellen (1985) focuses on the possibility that near rationality can account for monetary disequilibrium. This analysis directly confronts the problem that the postulate of maximization is inconsistent with an effect of monetary policy on real activity. It poses the questions of how much non-maximizing behaviour is necessary, and what form this behaviour must take, in order for the effects of monetary disturbances on real activity to have a realistic order of magnitude. Akerlof and Yellen show that minor deviations from maximization by a subset of producers, who individually suffer only second-order consequences, are sufficient to produce first-order macroeconomic effects.

These recent developments still leave us without a fully unified theoretical framework applicable to the analysis of macroeconomic fluctuations and to the analysis of resource allocation and income distribution. Apparently, economic theory in its present state has to rely on empirical regularities to identify the sets of questions for which either near rationality of full rationality are more useful 'as if' assumptions.

See Also

- ▶ Equilibrium: An Expectational Concept
- ▶ Monetary Equilibrium
- ▶ Money and General Equilibrium Theory

Bibliography

Akerlof, G., and J. Yellen. 1985. A near-rational model of the business cycle with wage and price inertia. *Quarterly Journal of Economics* 100(402, Supplement): 823–838.

Azariadis, C. 1978. Escalation clauses and the allocation of cyclical risks. *Journal of Economic Theory* 18(1): 119–155.

Barro, R.J. 1976. Rational expectations and the role of monetary policy. *Journal of Monetary Economics* 2(1): 1–32. Reprinted as ch. 3 in R.J. Barro, *Money, expectations, and business cycles*. New York:

Academic Press, 1981; also reprinted in *Rational expectations and econometric practice*, ed. R.E. Lucas Jr. and T.J. Sargent. Minneapolis: University of Minnesota Press, 1981.
Barro, R.J. 1981a. The equilibrium approach to business cycles, Ch. 2. In *Money, expectations, and business cycles*. New York: Academic Press.
Barro, R.J. 1981b. Unanticipated money growth and economic activity in the United States, Ch. 5. In *Money, expectations, and business cycles*. New York: Academic Press.
Barro, R.J., and Z. Hercovitz. 1980. Money stock revisions and unanticipated money growth. *Journal of Monetary Economics* 6(2): 257–267.
Boschen, J., and H.I. Grossman. 1982. Tests of equilibrium macroeconomics using contemporaneous monetary data. *Journal of Monetary Economics* 10(3): 309–333.
Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85(1): 191–205. Reprinted in *Rational expectations and econometric practice*, ed. R.E. Lucas Jr. and T.J. Sargent. Minneapolis: University of Minnesota Press, 1981.
Grossman, H.I. 1981. Incomplete information, risk shifting, and employment fluctuations. *Review of Economic Studies* 48(2): 189–197.
Grossman, H.I. 1983. The natural-rate hypothesis, the rational-expectations hypothesis, and the remarkable survival of non-market-clearing assumptions. *Carnegie-Rochester Conference Series on Public Policy* 19: 225–245.
Grossman, H.I. and Haraf, W.S. 1985. *Shunto, rational expectations, and output growth in Japan*. NBER Working Paper No. 1144, revised July 1985.
King, R.G. 1981. Monetary information and monetary neutrality. *Journal of Monetary Economics* 7(2): 195–206.
King, R.G., and C.I. Plosser. 1984. Money, credit, and prices in a real business cycle. *American Economic Review* 74(3): 363–380.
Lucas Jr., R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4(2):103–124. Reprinted in R.E. Lucas Jr., *Studies in business cycle theory*. Cambridge, MA: MIT Press, 1981.
Lucas Jr., R.E. 1973. Some international evidence on output-inflation tradeoffs. *American Economic review* 63(3): 326–334. Reprinted in R.E. Lucas Jr., *Studies in business-cycle theory*. Cambridge, MA: MIT Press, 1981.
Lucas Jr., R.E. 1985. *Models of business cycles*, Yrjö Jahnsson Lectures. Helsinki.
McCallum, B.T. 1986. On 'real' and 'sticky-price' theories of the business cycle. *Journal of Money, Credit and Banking* 17.
Schultze, C.L. 1985. Microeconomic efficiency and nominal wage stickiness. *American Economic Review* 75(1): 1–15.
Yeager, L. 1986. The significance of monetary disequilibrium. *Cato Journal* 6.

Monetary Economics, History of

Robert W. Dimand

Abstract

Throughout its long history, monetary economics has been concerned with the role of money in exchange, with what determines the purchasing power of money, and with the effects of changes in the purchasing power of money on interest rates and real economic activity.

Keywords

Aggregate demand; Allais, M; Aristotle; Bagehot, W; Banking crises; Banking school; Barter; Bentham, J; Bimetallism; Bullionist controversy; Cantillon, R; Central banking; Convertibility; Credit cycles; Currency school; Debasement of currency; Distributed lags; Effective demand; Equation of exchange; Exchange; Fiat money; Fisher, I; Forced saving; Free banking; Friedman, M; Fullarton, J; General equilibrium approach to monetary theory; Gold standard; Hawtrey, R. G; Hicks, J. R; Hume, D; Index numbers; Inflation; Inflation; Interest rate limits; Involuntary unemployment; Jevons, W. S; Keynes, J. M; Malthus, T. R; Markowitz, H. M; Marshall, A; Medium of exchange; Mercantilism; Mill. J. S; Monetarism; Monetary approach to the balance of payments; Monetary economics; Monetary policy rules; Money; Money illusion; Money supply; Mun, T; Natural rate and market rate of interest; Neutrality of money; Nominal interest rates; Open market operations; Overlapping generations models; Pigou, A. C; Plato; Portfolio choice; Price revolution; Quantity theory of money; Real balances; Real bills doctrine; Ricardo, D; Robertson, D. H; Say's equality; Say's identity; Say's law; Sismondi, J. C. L. S. de; Smith, A; Specie-flow mechanism; Stabilization; Stockholm school; Symmetallism; Taylor rule; Term structure of interest rates; Thornton, H; time preference;

Tobin, J; Tooke, T; Turgot, A.R. J; Uncovered interest parity; Unemployment; Veil of money; Velocity of circulation]; Walras, L; Walras's, Law; Wicksell, J. G. K

JEL Classification
B2

Origins of Monetary Economics

As with so much else in the Western tradition, theorizing about the role of money can be traced back to Plato and Aristotle in the fourth century BCE, although they may have drawn on pre-Socratic philosophers whose works survive, if at all, only in fragments. In his *Republic* (1974), Plato remarked that money was a symbol devised to make exchange easier. He disapproved of gold and silver as money, preferring a currency that would have value only internally, not in external commerce. The analysis in Aristotle's *Nicomachean Ethics* (1996) and *Politics* (1984) of what constitutes just exchange led Aristotle to a more systematic discussion of a medium of exchange. His account of the functions of money, and of the properties that suit a commodity such as gold or silver to be the medium of exchange, as well as his use of the myth of Midas to distinguish between gold and wealth, influenced comparable presentations by Nicolas Oresme in about 1360 (Oresme et al. 1989), Adam Smith (1776), and, through Smith, any number of nineteenth-century textbooks (see Menger 1892; Monroe 1923). Barter might be the most basic form of exchange, but it involves accepting goods one does not wish to consume in order to make a further exchange for what is desired. Aristotle noted the convenience of a generally accepted medium of exchange in reducing the number of transactions required. He saw the convenience of stating prices in terms of the medium of exchange, and that, if a commodity is to serve as a medium of exchange, it must also be a store of value, retaining purchasing power between being received and being spent (but he did not mention the function of money as a standard of deferred payment). Precious metals provided a suitable medium of exchange because of being homogenous, divisible, portable, and sufficiently scarce to a have a high value relative to their weight, although that value could change. Unlike Plato, Aristotle viewed the weight and purity of the precious metals as the source of the purchasing power of money, with coinage just saving the inconvenience of having to weigh and assay the metals at every transaction.

The quantity theory of money, described by David Laidler (1991b) as 'always and everywhere controversial' and by Mark Blaug as 'the oldest surviving theory in economics' (Blaug et al. 1995), holds that the price level (the inverse of the purchasing power of money) depends on how large the stock of money is compared with the demand for real money balances, with the direction of causation running from money to prices (Hegeland 1951). The quantity theory originated in the sixteenth century, when Martin de Azpilcueta Navarro in Salamanca in 1556 and Jean Bodin in France in 1568 identified the inflow of silver from the Spanish colonies of Mexico and Upper Peru as the cause of the rise in prices and depreciation of silver throughout Europe, a phenomenon now known as the 'price revolution' (Grice-Hutchinson 1952; O'Brien 2000). In contrast to the recognition by Navarro and Bodin of the inverse relationship between the quantity of the precious metals and their purchasing power, contemporaries such as the Seigneur de Malestroit had attributed rising prices of commodities to the debasement of various national coinages. The astronomer Copernicus had remarked earlier that money usually depreciates when it is too abundant (Grice-Hutchinson 1952, p. 34), but Navarro and Bodin went beyond such passing insights to formulate a theory they could use to explain the observed trend of commodity prices. Later research has shown that the sixteenth century quadrupling of prices was also due in part to the growing output of central European silver mines and to an increase in the velocity of circulation of money as systems of payment and communication evolved, notably the use of bills of exchange.

Mercantilists also took note of the inflow of precious metals from Spain's conquest in the New World, viewing this gold and silver as the 'sinews of war' with which Spain could pay armies in Europe. Although both alchemy and seizure of the Spanish treasure fleet were attempted (the physicist Isaac Newton was both Master of the Royal Mint and an avid alchemist), mercantilists such as Thomas Mun advocated interventionist government policies to achieve a surplus of exports over imports as the way to bring gold and silver into a country that lacked its own mines. Mercantilists held that increased circulation of gold and silver in a country would both increase national power and stimulate real economic activity (Viner 1937; Vickers 1959). Isaac Gervaise (1720), Richard Cantillon (2001, written c. 1730 and published posthumously in 1755), and, most fully and forcefully, David Hume (1752) used the quantity theory of money to develop the specie-flow mechanism of international payments adjustment that rendered such mercantilist schemes futile. An increase in gold and silver circulating in a country, whether due to colonial conquests, discovery of new mines, or a trade surplus engineered by tariffs on imports and bounties on exports, would increase spending. Although Hume recognized that one immediate, temporary effect of such increased spending would be to stimulate production (see Humphrey 1993) in due course prices and wages would rise, making domestic goods more expensive in relation to foreign goods. This would reduce exports and increase imports, eliminating the trade surplus, so that the only lasting result would be the misallocation of resources caused by tariffs, bounties and quotas. For Adam Smith (1776), a small open economy such as that of Scotland took prices under the gold standard as given by the world market, so the balance of payments adjustment would take place without any change in the relative price of foreign and domestic goods. An excess supply of money in a country would directly cause more imports and more exportable goods to be purchased domestically (and the contrary in a country with an excess demand for money) unless the world's supply of monetary metal was distributed across countries in proportion to their demand for money. Humphrey (1993) and Laidler (2003, ch. 1) show that Smith's analysis bore a closer resemblance than that of Hume to the modern monetary approach to the balance of payments.

From Aristotle and the Bible onwards, payment of interest on loans had been condemned as usury on the grounds that it was unnatural for gold ('barren metal') to breed and that interest violated justice (exchange of equal values), as the amount of money repaid exceeded the initial loan. Cantillon, Hume, A.R.J. Turgot, and Jeremy Bentham argued for the legitimacy of an interest rate set by market forces of supply and demand, with Turgot invoking time preference to point out that the amount of money lent and the larger amount of money repaid represented the same present value. Contrary to his general stand against government intervention, Adam Smith (1776) endorsed legal limits on interest to prevent high-risk lending for speculation and reckless consumption, and was rebuked for inconsistency by the young Bentham (West 1997).

Monetary Controversies in Classical Economics

Monetary theory was advanced by two British debates, the Bullionist Controversy, which surrounded the suspension of the convertibility of Bank of England notes into gold from 1797 to 1821, and the clash between the Banking School and the Currency School in the 1840s leading up to and following the Bank Act revision that separated and regulated the Bank of England's Issue Department (whose liabilities were bank notes, with gold held in reserve) and Banking Department (whose liabilities were deposits, with Bank of England notes held in reserve). During the suspension of convertibility during the Napoleonic Wars, Henry Thornton (1802) and, from 1809 onwards, David Ricardo (1810) argued the high price of bullion and foreign exchange showed that the Bank of England had engaged in over-issue of bank notes, raising commodity prices and depreciating the pound sterling (Fetter 1965; Marcuzzo and Rosselli 1991). Christiernin

(1761) had made a similar argument in Sweden, but appears not to have been known in Britain. Thornton was the leading figure on a House of Commons Select Committee on the High Price of Gold Bullion in 1810 that adopted this view in the Bullion Report, but the directors of the Bank of England persuaded the full House not to act on the committee's report. The directors, invoking the authority of Adam Smith, held that they could not have been guilty of any inflationary overissue of notes beyond what the needs of trade required as long as they issued notes only by discounting bills of exchange created by genuine commercial transactions, rather than financial speculation. This version of the real bills doctrine ignored Smith's assumption that bank notes were convertible into gold upon demand, so that any increase in the quantity of notes sufficient to depress their value below their gold par would cause the excess notes to be redeemed. Without convertibility as a constraint on overissue, the demand for bills would be unbounded as long as the discount rate was less than the prevailing rate of profit. The distinction between real and fictitious bills also failed to recognize that the length of time a bill was discounted need not correspond to the length of time goods were in process (Mints 1945; Laidler 2003; Davis 2005).

The depression that accompanied the end of the Napoleonic Wars and Britain's subsequent return to the gold standard stimulated a debate over the possibility of a general glut of commodities. Thomas Robert Malthus and J.C.L. Simonde de Sismondi attributed the depression to an insufficiency of effective demand. Malthus's argument was acclaimed by John Maynard Keynes a century later, although, unlike Keynes, Malthus did not distinguish between a decision to save and a decision to invest (see Keynes's 1933 essay on Malthus). Ricardo and Jean-Baptiste Say upheld Say's (or James Mill's) Law of Markets, denying the possibility of a general glut of commodities or an insufficiency of aggregate effective demand, since a commodity was offered for sale only with the intention of acquiring the means to purchase some other commodity, not with intent to hoard money, which is only a medium of exchange (Say was not quite as unambiguous as James Mill). Ricardo and Say recognized that unemployment would occur during the adjustment to a major change in the mix of commodities demanded, as the end of the Napoleonic Wars curtailed military and naval spending and as the purchasing power of money changed: Ricardo was prepared to accept restoration of gold convertibility at the depreciated parity, to avoid the price deflation associated with going back to the pre-war parity, and Say endorsed public works to employ those who would otherwise be jobless during the transition period. But, according to Ricardo, Say and James Mill, such distress resulted from a temporary mismatch between the mix of commodities produced and those demanded, with excess supply in some markets and excess demand in others, not from generalized excess supply.

Throughout the nineteenth century, classical economists such as John Stuart Mill struggled to formulate an acceptable version of the law of markets that would be stronger than what Oskar Lange later labelled Say's Equality but weaker than what Lange called Say's Identity (Corry 1962; Sowell 1972; Baumol 1977, 1999; Davis 2005). Say's equality, which held that at equilibrium prices the value of excess demand sums to zero across all markets except that for money, is a trivial implication of the market-clearing equilibrium condition that at market-clearing prices supply equals demand in each market. Say's identity, which held that at any prices the value of excess demand always sums to zero across all markets except money, implies (when combined with the summation of individual budget constraints) that money demand always equals the money supply at any prices, which leaves the absolute price level (the inverse of the purchasing power of money) indeterminate. In the 1870s, Leon Walras reformulated Say's Law as what Lange termed Walras's Law: the value of aggregate excess demand summed over all markets (including money) is identically zero, from the summation of individual budget constraints (the net value of each individual's transactions is at most zero, since people must pay for their purchases) plus local non-satiation (so that no one is willing to throw away purchasing power). Robert Clower

(1984), seeking to understand Keynes's rejection of Say's Law of Markets, argued that Walras's Law only applies to notional demands, not to quantity-constrained effective demands when markets do not clear (in Keynes's case, the labour market): if workers cannot sell all the labour they wish at the prevailing wage rate, then the quantity of labour they cannot sell multiplied by the wage rate that they would have received should not be included in their budget constraint for demanding goods.

Currency School adherents (for example, J.R. McCulloch, G.W. Norman and Lord Overstone), whose ideas shaped Sir Robert Peel's Bank Act of 1844, urged that, beyond maintaining convertibility, the Bank of England should conduct its operations so that a mixed metallic and paper currency would fluctuate in the same way that a purely metallic currency would. Building on Ricardo's presentation of the quantity theory of money and the price specie-flow mechanism, the Currency School wished the central bank to follow a stabilizing policy that would prevent gold outflows, rather than waiting for such international cash drains to bring about adjustment. The Currency School attributed the banking crises of 1825, 1832 and 1836–1837 to monetary mismanagement by the Bank of England, which could have regulated the volume of coin and notes in circulation so as to stabilize prices. In contrast, Banking School writers such as Thomas Tooke and John Fullarton, drawing on Thornton, emphasized the endogeneity of the total volume of credit (financial instruments convertible into gold), of which bank notes were only a small part (Fullarton 1836, 1845; Fetter 1965; Arnon 1991; Cassidy 1998; Skaggs 1999). Karl Marx also held that the volume of money adjusted to satisfy the equation of exchange (de Brunhoff 1976). Elements of both Currency School and Banking School positions appeared in the writings of John Stuart Mill. The Banking School thought that the volume of credit was as likely to respond to changes in prices as to cause them, and so did not share the Currency School view of the banking system as the initiator of credit cycles. The Banking School prescription was for the Bank of England to hold a bullion reserve large enough to ride out temporary disturbances in credit and international payments. While the Currency and Banking Schools differed on the appropriate policy for a central bank, another group of writers, including Henry Dunning Macleod (1855), James Wilson of *The Economist* and Jean-Gustave Courcelle-Seneuil, opposed having a central bank with a legally protected dominant position and special privileges. Instead, they advocated a system of free banking, with the market valuing the notes of competing banks, a proposal revived by Vera Smith (1936) and later by Friedrich Hayek (1976), who had been her dissertation adviser. Walter Bagehot's *Lombard Street* (1873) established the monetary orthodoxy, emerging from the Currency School–Banking School debates, on how the central bank should manage the discount rate to maintain convertibility and its role as a lender of last resort to preserve the liquidity of the banking system, rather than simply acting in the interests of its shareholders.

The Golden Age of the Quantity Theory

In studies collected posthumously in Jevons (1884), William Stanley Jevons used index numbers, with equal weights on different commodities, to show the rise in prices following the gold rushes in California in 1849 and Australia in 1851, as did John Elliot Cairnes. Commodity prices tended downwards from 1873 to 1896 as the world's demand for real money balances grew faster than its money supply, a decline halted by the introduction of the cyanide process for extracting gold from low-grade ores and by gold discoveries in South Africa and the Klondike. Together with the return of the United States to gold convertibility of the dollar in 1873 after the issue of inconvertible greenbacks during the Civil War, this deflation contributed to bimetallist agitation that reached its peak in William Jennings Bryan's presidential campaign in 1896, in which Bryan spoke against 'crucifying mankind on a cross of gold'. The bimetallists argued that monetizing silver as well as gold would raise the price by increasing the quantity of money, and this would have lasting real benefits. This led hard-

money, classical economists such as J. Laurence Laughlin of the University of Chicago to associate the quantity theory of money with claims of long-run non-neutrality (Skaggs 1995). In place of the quantity theory, Laughlin (1903) derived the value of money from the convertibility into gold, whose value depended on its cost of production, a view which David Glasner (1985, 2000) shows had figured alongside the quantity theory in classical political economy. The quantity theorists David Kinley (1904), Edwin Kemmerer (1907) and Irving Fisher (with Harry G. Brown, *The Purchasing Power of Money*, 1911, in Fisher 1997, vol. 4) responded by seeking to show, contrary to Laughlin and his Chicago associates, that exogenous changes in the quantity of money explained the behaviour of prices (given the trend in money demand), and, contrary to the bimetallists, that money is neutral in the long run. These quantity theorists extended earlier statements of the equation of exchange by Simon Newcomb (to whom Fisher dedicated his 1911 book) and Sir John Lubbock. Fisher allowed currency (M) and bank deposits (M′) to have different velocities of circulation, restating the equation of exchange as $MV + M'V' = PT$, where T is an index of the volume of transactions and P is the price level. To use the equation of exchange to make the case that the changing money supply explained the observed movements of US prices (rather than just having the equation as a tautology defining the velocity of circulation) required independent measures of the velocity of circulation. To estimate V, Fisher persuaded 116 people at Yale (including 113 male undergraduates) to keep daily records of their spending and cash balances. For V′, the velocity of circulation of bank deposits, Fisher used linear interpolation between the estimates from two empirical studies by David Kinley counting all bank clearings in the United States for a day in 1896 (for the Comptroller of the Currency) and a day in 1910 (for the National Monetary Commission). From an Austrian perspective, Ludwig von Mises (1935) objected to the aggregative reasoning of the quantity theorists, arguing that an index number of the price level gives a distorted picture of how agents respond to prices.

Systematically developing earlier remarks by John Stuart Mill and Alfred Marshall and an article by Jacob de Haas, Irving Fisher argued in *Appreciation and Interest* (1896, in Fisher 1997, vol. 1) that that nominal interest is the sum of real interest and the expected rate of inflation, so that only unanticipated changes in the purchasing power of money change the real interest rate and redistribute wealth. Contrary to bimetallist claims, expected inflation or deflation would have no real effects. Fisher's 1896 analysis included uncovered interest arbitrage parity (the difference between nominal interest rates in two currencies is the expected rate of change of the exchange rate) and the expectations theory of the term structure of interest rates (variations in nominal interest on loans of different duration reflects expectations of the time-path of prices). But from *The Purchasing Power of Money* onwards, while continuing to insist on the long-run neutrality of money, Fisher argued that money was not neutral during transition periods (of up to 10 years), as nominal interest adjusted only slowly to monetary shocks, and that the 'so-called "business cycle"' was really a 'dance of the dollar'. While Ralph Hawtrey (1919) and Fisher advanced monetary theories of economic fluctuations, many economists in the late nineteenth and early twentieth centuries, from Jevons on sunspot cycles to Joseph Schumpeter on clusters of innovations, emphasized real shocks and truly periodic cycles of varying lengths such as Juglar, Kondratiev and Kitchin cycles. Fisher's article, 'A Statistical Relationship between Unemployment and Price Level Changes' (1926) correlated unemployment with a distributed lag of past price level changes and was reprinted in the *Journal of Political Economy* in 1973 as 'Lost and Found: I Discovered the Phillips Curve'. Fisher correlated nominal interest with a distributed lag of price changes (a version of adaptive expectations) to show the slow adjustment of nominal interest and inflation expectations (*The Theory of Interest*, 1930), resulting from what he termed *The Money Illusion* (the title of his 1928 book), the widespread tendency to think in nominal rather than real terms.

Bimetallism foundered on its insistence on fixing the relative price of gold and silver, at 15 or

16 ounces of silver per ounce of gold. As the relative market valuation changed, due to changing marginal costs of production or shifts in non-monetary demand for precious metals, one of the two metals would disappear from circulation and its coins be melted down. Alfred Marshall's (1887) suggestion of symmetallism, a unit of value consisting of a quantity of gold plus a quantity of silver (reprinted in Pigou 1925), was more practical, but did not seem so to bimetallists or the general public. Marshall's tentative proposal to peg the monetary value of a basket of two commodities instead of just one (gold) marked a step towards a monetary policy of targeting the price level (or its rate of change) rather than the exchange rate with gold. Like Jevons (1884), Marshall suggested voluntary indexation, with contracts made in terms of a 'standard unit of purchasing power', which Marshall argued would reduce cyclical fluctuations (Laidler 1991a, pp. 172–8). Irving Fisher and Senator Robert Owen attempted unsuccessfully to get such a price level target into the Federal Reserve Act of 1913. The Federal Reserve Act, influenced by J.L. Laughlin and his student H. Parker Willis, instead adopted a fixed price of gold and, inconsistent with that goal, a version of the real bills doctrine that the volume of currency and bank credit should vary pro-cyclically with the needs of trade. As Knut Wicksell (1915) and others objected, Fisher compromised his compensated dollar plan by disguising it as a version of the gold standard, with the gold weight of the dollar changed periodically to peg the dollar price of a basket of commodities, a system vulnerable to speculative attacks. By 1935, when Fisher endorsed open market operations under a floating exchange rate to achieve a price-level target, he had lost his audience.

While Fisher distinguished nominal and real interest rates, Knut Wicksell (1898, 1915) stressed the distinction between the market rate of interest, set by the banking system, and the natural rate of interest that would equilibrate desired investment and saving (Laidler 1991a; Humphrey 1993). As long as the market rate is less than the natural rate, entrepreneurs can profit by borrowing and investing, causing total spending to increase and prices to rise. Such a cumulative inflation would continue until the growth of loans and deposits and a drain of cash out of the banking system reduced the ratio of reserves to bank deposits, forcing banks to raise the market rate to restore their liquidity. Wicksell pointed out that in a cashless economy, with only bank money used for transactions and no reserves held by banks, there would be no such force to automatically halt a cumulative inflation or deflation, and stability would depend on deliberate action by the monetary authority to match the market rate to the changing natural rate. To explain observed price movements, Wicksell emphasized real shocks that changed the natural rate as initiating fluctuations. Wicksell's two-rate model greatly influenced the Stockholm School (Karin Kock, Erik Lindahl, Erik Lundberg, Gunnar Myrdal, Bertil Ohlin) and John Maynard Keynes's *Treatise on Money* (1930). Recent financial innovations, diminishing the role of money as a means of payment and as an asset, have renewed attention to Wicksell's analysis of a cashless economy in which the monetary authority pursues stabilization by setting the interest rate rather than the quantity of money. The title of Michael Woodford's (2003) *Interest and Prices* deliberately echoes the title of Wicksell's (1898) *Interest and Prices* and a change of emphasis from Don Patinkin's (1965) *Money, Interest and Prices*. The 'Taylor rule', the influential monetary policy rule proposed by John Taylor, amounts to an attempt to set the market rate of interest equal to a Wicksellian natural rate that changes over time and is not directly observable.

Cambridge Monetary Theory and the Keynesian Revolution

In his lectures at Cambridge, evidence to official inquiries (collected by Keynes after Marshall's death as Marshall 1926), and manuscripts from the 1870s that half a century later formed the basis of Marshall (1923), Alfred Marshall expounded the quantity theory of money in a version that emphasized that desired cash balances are proportional to nominal income, $M = kPY$ (see Robertson 1922; Marget 1938–1942; Eshag 1963;

Bridel 1987; Laidler 1999 on Cambridge monetary economics). The Cambridge coefficient k is the reciprocal of V, the income velocity of circulation of money in the equation of exchange, so that the two versions of the quantity theory are formally equivalent, although Marshall's disciples A.C. Pigou and J.M. Keynes claimed that Cambridge discussions of the determinants of k were more choice-theoretic and less mechanical than Fisher's discussion of the determinants of velocity. Related contributions emerged from both traditions: Fisher was the first to correctly state the marginal opportunity cost of holding real money balances (1930), Keynes the first to explicitly write money demand as a function of income and nominal interest (*General Theory*, 1936). Writing in a time of floating exchange rates and Continental European hyperinflations after the First World War, the young Keynes, in *A Tract on Monetary Reform* (1923), extended Marshall's monetary economics to analyse inflation as a tax on holding money and government bonds, the social costs of inflation (both distortions from incorrectly anticipated inflation and higher transactions costs as expected inflation reduces the demand for real money balances), and covered interest arbitrage parity (the spread between spot and forward exchange rates is the difference between nominal interest in two currencies). Keynes opposed Britain's return to the gold standard at the pre-war parity in 1925 as entailing domestic deflation and, until wages declined, unemployment. Keynes's position recalled Ricardo's preference for restoring convertibility as a depreciated parity after the Napoleonic Wars. D.H. Robertson (1926), deeply Marshallian although a student of Keynes and Pigou rather than directly of Marshall, examined the effect of price level changes on saving and investment, notably how an increase in the price level causes forced saving ('induced lacking') to restore real money balances (Laidler 1999).

Reflecting on Britain's stagnation after the return to gold and on the worldwide Great Depression of the 1930s, Keynes's *General Theory of Employment, Interest and Money* (1936) denied the automatic restoration of full employment in a monetary economy after a negative demand shock. Keynes lumped together economists from Ricardo to Marshall and Pigou as 'classical' economists who accepted Say's Law (summarized by Keynes as 'supply creates its own demand'). Keynes subsequently clarified that he did not regard Fisher, Hawtrey, Robertson or Wicksell's Swedish followers as classical (but he did think that Wicksell himself was trying to be classical), and, as Ellis (1934) showed, German monetary theorists such as Joseph Schumpeter and L. Albert Hahn were far from classical about the real effects of an expansion of the banking system. In contrast to von Mises (1935) and Hayek (1931), who viewed depressions as necessary corrections of earlier overinvestment, Keynes held that depressions were calamities that the government and monetary authority could overcome by increasing aggregate demand, rather than relying on wage and price deflation to restore full employment. Keynes considered it crucial that wage bargains are made in money terms, so that workers concerned about relative wages might accept a price level increase to clear the labour market while quite rationally opposing money wage cuts as staggered contracts came up for renegotiation (1936, ch. 2). Wage cuts, and the associated deflation of prices, would increase demand for real money balances, exerting a contractionary effect on aggregate demand (1936, ch. 19). Keynes identified volatile private investment, resulting from fundamental uncertainty about future profitability, as the source of economic fluctuations, and, like the generations of Keynesian, New Keynesian and Post Keynesian economists after him, saw a need for management of aggregate demand to stabilize the economy.

The Revival of the Quantity Theory of Money

While Keynes was arguing the case for stabilization policy, Henry Simons of the University of Chicago made the case for rules rather than discretion in monetary policy (Simons 1936). Keynes saw a role for government to counteract the instability resulting from volatile private spending, but Chicago quantity theorists (later called monetarists) such as Simons (1936) and Milton Friedman and his students (Friedman 1956) blamed volatile,

unpredictable monetary policy for economic instability. Keynesians invoked the Great Depression of the 1930s as demonstrating the need for government stabilization of an unstable private sector in a monetary economy, but Friedman and Anna J. Schwartz (1963) blamed the depression on a misguided Federal Reserve system that permitted a 'great contraction' of the money supply. Misled by the real bills doctrine, the Federal Reserve Board had not paid sufficient attention to the quantity of money. Where Keynes had emphasized the fundamental uncertainty underlying long-period expectations of profitability, Friedman (like Fisher) stressed the endogeneity of expectations of inflation: people cannot be fooled indefinitely by inflation into working more for a lower real wage that they think they are getting, because they will learn from experience (see Friedman and his critics in Gordon 1974). Keynes worried about involuntary unemployment – an excess supply of labour because the labour market did not clear – while Friedman held that at any correctly anticipated inflation rate unemployment would be at its natural rate, reflecting voluntary investment in search and consumption of leisure. Friedman claimed in 1956 to be following a Chicago oral tradition of monetary theory taught by Frank Knight, Jacob Viner, Henry Simons and Lloyd Mints that had replaced J. Laurence Laughlin's opposition to the quantity theory. Don Patinkin (1981) and David Laidler (2003), who both held Chicago Ph.D.s, argued that Friedman overstated the purely Chicago sources of his monetarism: Friedman's teachers had taught the works of non-Chicago quantity theorists such as Fisher as well as Keynes's earlier Marshallian *Tract on Monetary Reform* (1923) and his Wicksell-influenced *Treatise on Money* (1930). Friedman took a course in which the main textbook was Keynes's *Treatise*, which Keynes's detailed and extensive contribution to monetary analysis. Fisher had advocated a monetary policy rule (a price level target, rather than the constant of money growth proposed by Friedman), while Keynes's *Tract* was as attentive as any Chicago monetarist to the social costs of inflation. A key element of Friedman's monetarism, money demand as a function of a small list of variables, had first appeared in Keynes's *General Theory*. There were also parallel, independent revivals of the quantity theory of money far from Chicago, such as that associated with Marius Holtrop, long-time president of the Netherlands central bank (De Jong 1973).

Integrating the Theory of Money into General Economic Theory

Rationalizing the use of money has been a problem in the development of general equilibrium theory: if markets are complete, or all debts will be repaid with certainty, there is no need for a particular asset to be singled out as a generally accepted means of payment. Irving Fisher's 1892 dissertation introduced general equilibrium analysis in North America, but he did not integrate his later monetary economics into a general equilibrium framework. Leon Walras, the founder of general equilibrium theory, wrote on the theory of money (for example, Walras 1886), starting with the equation of exchange and later discussing desired cash balances, *encaisse désirée*, but simply assumed that monetary exchange is superior to barter, rather than demonstrating that the use of money reduces transactions costs: 'In Walras's economy, agents hold money not out of choice but of a technological necessity' (Bridel 1997, p. 119; see also Patinkin 1965, pp. 531–72). In Walras's analysis, prices were stated in terms of a particular commodity, the *numéraire*, but it was not clear why transactions should use that commodity. The idea that money is only a veil over the real side of the economy long predates the introduction of the term 'veil of money' in English by Dennis Robertson (1922) and of 'neutrality of money' by Hayek (1931): (see Pigou 1949; Patinkin and Steiger 1989). Don Patinkin (1965) argued that a long list of classical and neoclassical economists postulated, at least implicitly, an invalid dichotomy between the real and nominal sides of the economy, in which an equi-proportional change in all money prices (so that no relative prices changed) would not affect the excess demands for commodities. Such a dichotomy would exclude the real balance effect that would bring the general price level to equilibrium. The valid dichotomy would hold that an

equi-proportional change in all money prices, the quantity of money, and any exogenous nominal variables (such as quantities of government bonds) would have no real effects.

John Hicks (1935) set the agenda for much later work integrating the theory of money into the more general theory of value, seeking choice-theoretic explanations of why fiat money, not backed by convertibility into a commodity such as gold or silver, has a positive purchasing power, and why people choose to hold part of their wealth in money (either non-interest-bearing high-powered money or highly liquid close substitutes paying low rates of interest) rather than in alternative assets that pay a higher rate of return. Following Hicks's argument for treating the decision to hold money as part of the allocation of wealth across a portfolio of assets, James Tobin (1958) introduced money as a riskless asset (at least in nominal terms) into Harry Markowitz's theory of portfolio choice. Risk-averse individuals would divide their wealth between money (zero return, zero risk) and a portfolio of risky assets with positive expected return. Each investor would combine risky assets in the same proportions, differing from other investors only in the fraction of wealth held in the riskless asset. If returns were normally distributed or investors had quadratic loss functions, this portfolio choice could be conveniently captured by a two-dimensional diagram (the mean and standard distribution of portfolio returns), and if investors had constant relative risk aversion, the share of wealth held in each asset (including money) would be independent of the level of wealth (see Tobin 1958, 1969; Tobin and Golub 1998). However, money is a risky asset in real terms, as its purchasing power may be eroded by inflation, and is dominated in rate of return by such short-term, highly liquid assets as Treasury bills, which, like money, have no default risk. While Treasury bills have some nominal risk, since a rise in nominal interest would lower their market price, this risk is limited by the short maturity of the bills. Tobin (1969) extended his portfolio approach to a 'general equilibrium approach to monetary theory' that treated money as one of a range of imperfectly substitutable assets whose rates of return are determined simultaneously, with an adding-up constraint that asset demands sum to total wealth, but without assuming continuous clearing of non-financial markets (Tobin 1971; Tobin and Golub 1998).

Another approach to a choice-theoretic explanation of demand for fiat money assumes that money must be used as a means of payment and that it is costly to trade between money and interest-bearing assets, so that individuals trade off the interest forgone by holding money against the transaction costs (including the value of one's time spent going to the bank) incurred by having to liquidate interest-bearing assets when having to make payments. Maurice Allais in 1947, William Baumol in 1952, and James Tobin in 1956 independently derived the square-root rule for this inventory approach to the transactions demand for money by minimizing the total cost of cash management, forgone interest plus transactions costs (see Allais 1947, pp. 238–41; Tobin and Golub 1998), unaware that Francis Ysidro Edgeworth (1888), followed by Wicksell (1898, pp. 57–8), had derived a similar square-root rule for the demand for reserves by banks given randomness in withdrawals of deposits.

Another explanation for a positive value of fiat money is provided by overlapping generations (OLG) models, pioneered independently by Allais (1947) and Paul Samuelson (1958). In OLG models, agents live for two periods, but produce consumption goods only when young. The young trade goods to the old in return for money in anticipation of being able to exchange that money for goods in the next period when they themselves are old. Such models explain the existence of positive-valued fiat money on the assumption that no other assets exist. Other efforts to provide microeconomic foundations for fiat money emphasize monitoring costs and default risks, so that liabilities of a single, more easily monitored monetary authority are less risky than private promissory notes and therefore more acceptable as means of payments.

The long history of monetary economics reveals several recurring issues: why fiat money has value, how the real and monetary sides of the economy are related, whether a central bank should follow a rule (and if so which rule) or

have discretion (or whether a central bank should even exist), is the lender of last resort function consistent with a policy rule, whether money has a special role or is just one of many assets and forms of credit, how should monetary exchange be incorporated in the general theory of value. Monetary analysis has also been focused and stimulated by external events and current policy issues: the 'price revolution' of the sixteenth century, the high price of bullion while the convertibility of Bank of England notes was suspended during the Napoleonic Wars, the Bank of England's charter coming up for renewal in 1844 after several banking crises, the decline in the purchasing power of gold following the California and Australian gold rushes and its appreciation from 1873 to 1896, the Continental European hyperinflations after the First World War, Britain's return to the gold exchange standard at the pre-war parity in 1925, and the Great Depression.

See Also

- ▶ Banking School, Currency School, Free Banking School
- ▶ Bullionist Controversies (Empirical Evidence)
- ▶ Equation of Exchange
- ▶ Natural Rate and Market Rate of Interest
- ▶ Quantity Theory of Money
- ▶ Real Bills Doctrine Versus the Quantity Theory

Bibliography

Allais, M. 1947. *Ėconomie et intérêt*. Paris: Librairie des Publications Officieles.

Aristotle. 1984. *The politics*. Trans. Carnes Lord. Chicago: University of Chicago Press.

Aristotle. 1996. *The Nicomachean ethics*. Trans. Harris Rackham Ware, Herts: Wordsworth Editions.

Arnon, A. 1991. *Thomas Tooke, pioneer of monetary theory*. Aldershot/Brookfield: Edward Elgar.

Bagehot, W. 1873. *Lombard street*. In Bagehot (1974–86).

Bagehot, W. 1974–1986. *The collected works of Walter Bagehot*. London: The Economist.

Baumol, W.J. 1977. Say's (at least) eight laws, or what say and James Mill may really have meant. *Economica* NS 44, 145–162.

Baumol, W.J. 1999. Retrospectives: Say's law. *Journal of Economic Perspectives* 13: 195–204.

Blaug, M., W. Eltis, D. O'Brien, D. Patinkin, R. Skidelsky, and G.E. Wood. 1995. *The quantity theory of money from Locke to Keynes and Friedman*. Aldershot/Brookfield: Edward Elgar.

Bridel, P. 1987. *Cambridge monetary thought: The development of saving-investment analysis*. Basingstoke: Macmillan.

Bridel, P. 1997. *Money and general equilibrium theory: From Walras to Pareto (1870–1923)*. Cheltenham: Edward Elgar.

Cantillon, R. 2001. *Essay on the nature of commerce in general*. Trans. H. Higgs. New Brunswick: Transaction.

Cassidy, M. 1998. The development of John Fullarton's monetary thought. *European Journal of the History of Economic Thought* 5: 509–536.

Christiernin, P.N. 1761. *Lectures on the high price of foreign exchange in Sweden*. Trans. In, *The Swedish bullionist controversy*, ed. R.V. Eagly. Philadelphia: American Philosophical Society, 1967.

Clower, R.W. 1984. *Money and markets: Essays by Robert W. Clower*, ed. Donald A. Walker. Cambridge: Cambridge University Press.

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

Davis, T. 2005. *Ricardo's macroeconomics: Money, trade cycles and growth*. Cambridge: Cambridge University Press.

de Boyer, J. 2003. *La pensée monétaire: Histoire et analyse*. Paris: Ėditions Les Solos.

de Brunhoff, S. 1976. *Marx on money*. Trans. M. Goldbloom. New York: Urizen.

De Jong, F.J. 1973. *Developments of monetary theory in the Netherlands*. Rotterdam: Rotterdam University Press.

Edgeworth, F.Y. 1888. Mathematical theory of banking. *Journal of the Royal Statistical Society* 51: 113–127.

Ellis, H.S. 1934. *German monetary theory 1905–1933*. Cambridge, MA: Harvard University Press.

Eshag, E. 1963. *From Marshall to Keynes: An essay on the monetary theory of the Cambridge School*. Oxford: Basil Blackwell.

Fetter, F.W. 1965. *The development of British monetary orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.

Fisher, I. 1892. *Mathematical investigations in the theory and value of prices*. New York: Macmillan In Fisher (1997), vol. 1.

Fisher, I. 1896. *Appreciation and interest*. In Fisher (1997), vol. 1.

Fisher, I. 1926. A statistical relationship between unemployment and price level changes. *International Labour Review* 13, 785–792. Repr. 1973 as Lost and found: I discovered the Phillips curve. *Journal of Political Economy* 81, 496–502. Also in Fisher (1997), vol. 8.

Fisher, I. 1928. *The money illusion*. In Fisher (1997), vol. 8.

Fisher, I. 1930. *The theory of interest*. In Fisher (1997), vol. 9.

Fisher, I. 1997. *The works of Irving Fisher*, 14 vols. ed. W.J. Barber, assisted R.W. Dimand and K. Foster. London: Pickering & Chatto.
Fisher, I., and H. Brown. 1911. *The purchasing power of money*. In Fisher (1997), vol. 4.
Friedman, M. 1956. *Studies in the quantity theory of money*. Chicago: University of Chicago Press.
Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States 1867–1960*. Princeton: Princeton University Press for the NBER.
Fullarton, J. 1836. Response to a proposal for a bank of India. Repr. 1998 in *European Journal of the History of Economic Thought* 5, 480–508.
Fullarton, J. 1845. *Regulation of currencies of the Bank of England*. New York: Augustus M. Kelley, 1969.
Gervaise, I. 1720. *The system or theory of trade of the world*. London: J. Roberts; Repr. Baltimore: Johns Hopkins Press, 1954.
Glasner, D. 1985. A reinterpretation of classical monetary theory. *Southern Economic Journal* 52: 46–68.
Glasner, D. 2000. Classical monetary theory and the quantity theory. *History of Political Economy* 32: 39–59.
Gonnard, R. 1936. *Histoire des doctrines monétaires, dans ses rapports avec l'histoire des monnaies*, 2 vols. Paris: Sirey.
Gordon, R.J. 1974. *Milton Friedman's monetary framework: A debate with his critics*. Chicago: University of Chicago Press.
Grice-Hutchinson, M. 1952. *The Salamanca school: Readings in Spanish monetary theory 1544–1605*. Oxford: Clarendon.
Guggenheim, T. 1989. *Preclassical monetary theories*. London/New York: Pinter.
Hawtrey, R.G. 1919. *Currency and credit*. London: Longmans, Green, 3rd edn. 1934.
Hayek, F.A. 1931. *Prices and production*. London: Routledge.
Hayek, F.A. 1976. *The denationalisation of money*. London: Institute of Economic Affairs.
Hegeland, H. 1951. *The quantity theory of money*. Göteborg/New York: Elanders Boktryckeri/Augustus M. Kelley, 1969.
Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* NS 2, 1–19.
Hume, D. 1752. *Writings on economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.
Humphrey, T.M. 1993. *Money, banking, and inflation: Essays in the history of economic thought*. Aldershot/Brookfield: Edward Elgar.
Jevons, W.S. 1875. *Money and the mechanism of exchange*. New York: D. Appleton, 1897.
Jevons, W.S. 1884. *Investigations in credit and prices*, ed. H.S. Foxwell. London: Macmillan.
Kemmerer, E.W. 1907. *Money and credit instruments in their relation to general prices*. New York: Henry Holt.
Keynes, J.M. 1923. *A tract on monetary reform*. In Keynes (1971–1989), vol. 4. Keynes, J.M. 1930. *Treatise on money*. In Keynes (1971–1989), vols 5 and 6.
Keynes, J.M. 1933. Robert Malthus: The first of the Cambridge economists. In *Essays in biography*, ed. J.M. Keynes. London: Macmillan. Repr. in Keynes (1971–1989), vol. 9.
Keynes, J.M. 1936. *General theory of employment, interest and money*. In Keynes (1971–1989), vol. 7.
Keynes, J.M. 1971–1989. *Collected writings of John Maynard Keynes*, 30 vols, ed. D.E. Moggridge and E.A.G. Robinson. London/New York: Macmillan/Cambridge University Press, for the Royal Economic Society.
Kinley, D. 1904. *Money, a study of the theory of the medium of exchange*. New York: Macmillan.
Laidler, D. 1991a. *The golden age of the quantity theory*. Princeton: Princeton University Press.
Laidler, D. 1991b. The quantity is always and everywhere controversial – Why? *Economic Record* 67: 289–306.
Laidler, D. 1999. *Fabricating the Keynesian revolution*. Cambridge: Cambridge University Press.
Laidler, D. 2003. *Macroeconomics in retrospect: Selected essays*. Cheltenham: Edward Elgar.
Laughlin, J.L. 1903. *The principles of money*. New York: Scribner.
Lowry, S. Todd. 1987. *The archaeology of economic ideas: The classical Greek tradition*. Durham: Duke University Press.
Macleod, H.D. 1855. *The theory and practice of banking*, 5th ed. London: Longmans/Green, 1893.
Marcuzzo, M.C., and A. Rosselli. 1991. *Ricardo and the gold standard: The foundations of the international monetary order*. London: Macmillan.
Marget, A.W. 1938–1942. *The theory of prices: A re-examination of the central problems of a monetary theory*, 2 vols. New York: Augustus M. Kelley, 1966.
Marshall, A. 1887. Remedies for fluctuations in general prices. *Contemporary review*, reprinted in *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan, 1925.
Marshall, A. 1923. *Money, credit and commerce*. London: Macmillan.
Marshall, A. 1926. *Official papers*, ed. J.M. Keynes. London: Macmillan.
Menger, C. 1892. On the origin of money. *Economic Journal* 2: 239–255.
Mints, L. 1945. *A history of banking theory in Great Britain and the United States*, 5th ed. Chicago: University of Chicago Press, 1970.
Monroe, A.E. 1923. *Monetary theory before Adam Smith*. New York: Augustus M. Kelley, 1969.
O'Brien, D.P. 2000. Bodin's analysis of inflation. *History of Political Economy* 32: 267–292.
Oresme, N., B. de Sassoferrato, and J. Buridan. 1989. *Traité des monnaies et autres écrits monétaires du XIVe siècle*, ed. C. Dupuy. Trans. F. Chartrain. Lyon: La Manufacture.
Patinkin, D. 1965. *Money, interest and prices*, 2nd ed. New York: Harper & Row.
Patinkin, D. 1981. *Essays on and in the Chicago tradition*. Durham: Duke University Press.

Patinkin, D., and O. Steiger. 1989. In search of the 'Veil of Money' and the 'Neutrality of Money': A note on the origin of terms. *Scandinavian Journal of Economics* 91: 131–146.
Pigou, A.C. 1925. *Memorials of Alfred Marshall*. London: Macmillan.
Pigou, A.C. 1949. *The veil of money*. London: Macmillan.
Plato. 1974. *The republic*. Trans. G.M.A. Grube. Indianapolis: Hackett Publishing.
Ricardo, D. 1810. *The high price of bullion, a proof of the depreciation of bank notes*. London: John Murray, 4th edn 1811. Repr. in Ricardo (1951–1973), vol. 3.
Ricardo, D. 1951–1973. *Works and correspondence of David Ricardo*, 11 vols, ed. P. Sraffa and M.H. Dobb. Cambridge: Cambridge University Press.
Rist, C. 1938. *Histoire des doctrines relative au crédit et à la monnaie depuis John Law jusqu'à nos jours*. Paris: Sirey. Trans. as *History of monetary and credit theory from John Law to the Present Day*, New York: Augustus M. Kelley, 1966.
Robertson, D.H. 1922. *Money*. Cambridge: Cambridge Economic Handbooks.
Robertson, D.H. 1926. *Banking policy and the price level*. London: P.S. King.
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.
Simons, H.C. 1936. Rules versus authorities in monetary policy. *Journal of Political Economy* 44: 1–30.
Skaggs, N.T. 1995. The methodological roots of J. Laurence Laughlin's anti-quantity theory of money and prices. *Journal of the History of Economic Thought* 17: 1–20.
Skaggs, N.T. 1999. Changing views: Twentieth-century opinion on the banking school-currency school controversy. *History of Political Economy* 31: 361–391.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. New York: Random House, 1937.
Smith, V.C. 1936. *The rationale of central banking*. London: P.S. King.
Sowell, T. 1972. *Say's law: An historical analysis*. Princeton: Princeton University Press.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*, with an introduction by F.A. Hayek. London: George Allen & Unwin, 1939. New York: Augustus M. Kelley, 1965.
Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1: 15–29.
Tobin, J. 1971. *Essays in economics, vol. 1, macroeconomics*. Chicago: Markham.
Tobin, J., and S. Golub. 1998. *Money, credit, and capital*. New York: McGraw-Hill.
Vickers, D. 1959. *Studies in the theory of money 1690–1776*. New York: Chilton.
Viner, J. 1937. *Studies in the theory of international trade*. London: George Allen & Unwin.
Von Mises, L. 1935. *The theory of money and credit*. Trans. H. Batson. London: Cape.
Walker, D.A. 1984. *Money and markets: Selected essays of Robert Clower*. Cambridge: Cambridge University Press.
Walras, L. 1886. *Théorie de la monnaie*. Paris: Ėditions Larose et Forcel.
West, E.G. 1997. Adam Smith's support for money and banking regulation: A case of inconsistency. *Journal of Money, Credit and Banking* 29: 127–135.
Wicksell, K.G. 1898. *Interest and prices*. Trans. R.F. Kahn. London: Macmillan, 1936.
Wicksell, K.G. 1915. *Lectures on political economy*, vol. 2, *money*. Trans. E. Claasen. London: Routledge, 1935.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

Monetary Equilibrium

Otto Steiger

The concept of monetary equilibrium is the fundamental feature of the macroeconomic theory originally formulated by Knut Wicksell (1898, 1906) and corrected, clarified and improved in the 1930s by Erik Lindahl (1930, 1934 and 1939b) and Gunnar Myrdal (1932, 1933 and 1939). Wicksell's approach was the first attempt to link the analysis of *relative prices* with the analysis of *money prices* (Shackle 1945, p. 47).

In the Wicksell–Lindahl–Myrdal theoretical structure the idea of a monetary equilibrium – the term stemmed from Myrdal (1932, p. 193) – was designed to analyse the conditions for equality of certain relations in a monetary economy which guarantee *macroeconomic equilibrium*, with the emphasis on a *stable price level*, as well as the implications of their non-fulfilment, that is, the consequences of *monetary disequilibrium*. In this analysis the notion of monetary equilibrium served not only as a theoretical tool, but also as an operational goal for economic policy.

Although frequently confused with the concept of *monetary neutrality* or *neutral money*, it has to

be emphasized that the notion of monetary equilibrium is conceptually distinct from this idea. The doctrine of neutral money – which also originated from Wicksell (Hayek 1931) – aimed to indicate the conditions under which the tendencies towards equilibrium in a barter economy, i.e. the equilibrium of relative prices according to neoclassical value theory 'are to remain operative in a monetary economy' (Hayek 1933, p. 160; cf. Koopmans 1933, p. 228). The theory of monetary equilibrium did not relate to these conditions, but to conditions of an equilibrium which the proponents of neutral money never intended to explain, still less to being regarded as a norm for economic policy.

The starting point of Wicksell's investigation into the conditions of monetary equilibrium, first presented in *Interest and Prices* (1898) and restated in *Lectures on Political Economy. Vol. II: Money* (1906), was his critical analysis of the attempts of both the dominating theories of value and the quantity theory of money to explain the value of money. These attempts had resulted in (i) a dichotomy of economic theory with entirely different laws for the value of money and the value of commodities and (ii) a theory of money which was unable to explain its postulated proportionality between changes in the quantity of money and the price level as the inverse of the value of money.

With regard to the first point, Wicksell (1903, p. 486f) had no difficulty in explaining the failure of both classical and neoclassical value theory to integrate monetary theory because of the impossibility of treating money as a commodity like all other commodities; therefore, they had to rely on the quantity theory to explain the value of money. This theory however – Wicksell's second point – holds true only under the assumption of a constant velocity of circulation as in the extreme case of 'a pure cash system without credit' (1898, p. 59). With credit, the velocity of circulation becomes a variable, and it is impossible to prove satisfactory and exact relationship between the quantity of money and the price level.

To solve the complications arising from money given or received as credit, Wicksell made the 'assumption' of a *pure credit economy* (cf. Palander 1941). By this device the quantity of money was determined endogenously by the demand for money and, therefore, abandoned as a direct price-determining force – a feature also common to the development of Wicksell's theory by Lindahl and Myrdal. Thus, freed from the tyranny of the quantity of money, Wicksell had to look for other forces determining the value of money.

To reveal these forces, he replaced the relation of the quantity theory between the quantity of money and the price level by a theory of the relation between the *interest on money loans* and the price level, which he analysed in the framework of two approaches: (i) the relation of the money or loan rate of interest as determined on 'the money market' to the 'natural' or real rate of interest as determined by the physical marginal productivity of capital (later replaced by value productivity); and (ii) the relation of aggregate monetary demand for and supply of commodities linked in the same manner as demand for and supply of an individual commodity. In his analysis Wicksell connected both approaches by showing that in a closed, competitive economy with a pure credit system, a deviation between the loan rate and the real rate of interest, by means of credit expansion or contraction, will serve as an incentive for entrepreneurs to invest or disinvest leading to a shift in the relation between aggregate monetary demand and supply which, under the assumption of given output, must result in a rise or fall in all money prices that due to anticipations of their initial changes becomes indefinite – Wicksell's famous *cumulative process*.

It becomes clear from this analysis that the cumulative process describes a system where the movements in money prices set no forces in operation towards an equilibrium. Wicksell considered, therefore, the nature of this monetary equilibrium as fundamentally distinct from the equilibrium of relative prices with its inherent tendency towards stability. Once disturbed, monetary equilibrium could be restored, however, by means of a special equilibrium rate, the so-called *normal rate of interest* on loans. Wicksell thought that under the more realistic premise of a mixed cash/credit system the changes in money prices as

'the connecting link' (1898, p. 109) between the money market and the commodity market would force the monetary authority to establish this rate.

However, Wicksell's concept of the normal rate was far from being clear and precise because it implied, as first shown by Lindahl (1930; cf. 1939a), three different conditions for monetary equilibrium: (i) to equal the natural or real rate, (ii) to equalize expected investment and saving and (iii) to preserve a stable price level, primarily of consumption goods. In their development of Wicksell's analysis both Lindahl and Myrdal attacked the consistency of this triple condition leading more or less to an abandonment of the notion of the normal rate by Lindahl and its reformulation by Myrdal.

With regard to the first condition Lindahl rejected to regard the loan rate as 'normal', since the level of the real rate could not be determined independently of it. Lindahl's concept of the real rate was characterized, in contrast to Wicksell (1898) but – as he later had to concede (1939b, p. 261) – in accordance with Wicksell's 'prospective profit rate' (1906), not by physical but by exchange *value* productivity, i.e. he defined the real rate as 'the relation between anticipated future product values . . . and the values invested' (1930, p. 124; 1939a, p. 248). As the demand for investment and, thereby, its price is influenced by the loan rate, the real rate will always have a tendency to adjust to the former. Therefore, the real rate could only have a meaning as that level of the loan rate which secures equilibrium between the expected values of investment and saving – Wicksell's second condition.

However, even this level of the loan rate is not 'normal' in the sense that it represents a unique equilibrium rate, since a change in investment, due to *any* shift in the loan rate, will always be balanced by a subsequent variation in the distribution of income between borrowers and lenders via changes in the price level. Thus, the second condition is fulfilled for different loan rates associated with different changes in the price level, and Lindahl abandoned, therefore, the concept of the normal rate in Wicksell's third condition for the notion of the '*neutral* rate of interest', that is a loan rate which is neutral in relation to *expected* changes in the price level, not to its constancy. However, as Lindahl realized that even this concept would still suffer from certain weaknesses, due to the difficulties of defining the price level with regard to different expectations as well as the many possible combinations of short and long term loan rates that are neutral in respect to the price level, he eventually decided not to employ the notion of a normal rate at all, confining himself to show 'that different interest levels . . . lead to different developments of the price level' (1930, p. 134; 1939a, p. 260).

Lindahl's position was immediately attacked by Myrdal in the original Swedish version of *Monetary Equilibrium* (1932), where the latter interpreted Lindahl's analysis as an attempt to get rid of the concept of monetary equilibrium. To prove this assertion Myrdal, like Lindahl, discussed Wicksell's three conditions – an analysis which led to a reconstruction of the concept of the normal rate and to 'a refutation of Lindahl's criticism of Wicksell' (Hansson 1981, p. 148).

With regard to Wicksell's first condition Myrdal tried to show that the real rate, contrary to Lindahl, could be treated as an independent entity determining the normal loan rate. In a response to Myrdal's criticism, Lindahl (1939b; cf. Hammarskjöld 1933) conceded that the different real rates, as visualized by an investment schedule in the capital market, could be considered indeed as independent of the current loan rate. However, it would be impossible from this schedule alone 'to single out any definite real rate as having a decisive influence on the loan rate', i.e. to determine the normal rate unless the corresponding saving schedule is known. Thus, Wicksell's first condition could be inferred only from his second. However, in the final English edition of *Monetary Equilibrium* Myrdal had already changed his mind (cf. Palander 1941). He now considered the determination of the first condition as being dependent on the second.

Myrdal's analysis of the first condition revealed the insight, that equality of the real and the loan rate as the condition, not for monetary equilibrium but for the determination of investment, could be used 'to explain *why* and *how* equilibrium is or is not maintained in the capital market' (Myrdal 1939, p. 87), that is, whether

Wicksell's second condition is or is not fulfilled which now became the sole criterion for monetary equilibrium and which Myrdal formulated in an *ex ante/ex post* framework (1939, ch. V; cf. 1932, pt. III; 1933, ch. V). This reformulation of the second condition was immediately accepted by Lindahl (1934; cf. 1939b, pp. 2d64–8) who, however, did *not* consider the equilibrium rate in the capital market as a *sufficient* condition for monetary equilibrium and developed instead a modified version of his concept of the 'neutral' rate as the normal rate.

In his investigation of Wicksell's third condition, Myrdal (1939, ch. VI; cf. 1932, pt. IV; 1933, ch. VI) concluded that monetary equilibrium is determined by the more fundamental first and second conditions, not by a stable price level. A uniform change in all money prices would neither change investment nor disturb equilibrium in the capital market, since monetary aggregates would vary in the same proportion. However, as price level changes are not uniform in reality where some money prices, like capital values, are highly flexible, while others, especially wages, are very sticky, the latter would 'act as a restraint on the price system' (1939, p. 134). Therefore, even if the third condition was deprived its significance for the determination of monetary equilibrium, it could be used as a *norm for monetary policy* aiming to restore a disrupted monetary equilibrium. As Myrdal emphasized, this does not mean a stabilization of the general price level but a *mitigation of the business cycle* brought about by an adaption of the flexible prices to the more sticky ones. This could be achieved by a stabilization of 'an index of those prices which are sticky in themselves' and which in practice would mean a stabilization of wages permitting capital values to move. For the case of monetary disequilibrium characterized by decreasing investment and increasing unemployment, Myrdal showed that such a depressive process could be stopped and reverted by a monetary policy supported by fiscal policy which first of all increases capital values to the level of the sticky wages, thereby preventing a fall of the latter which otherwise would aggrevate depression – as would a stabilization of capital values or any index of flexible prices. In spite of Myrdal's emphasis that 'the concept of monetary equilibrium has . . . central importance for the whole Wicksellian monetary theory' (1939, p. 30), both his and Lindahl's approaches are characterized by an obvious disinterest in equilibrium analysis and a preference for causistic disequilibrium analysis (Siven 1985) – a feature also common to the subsequent theories of the Stockholm School which, with the exception of Bent Hansen (1951, ch. 9), eventually discarded 'the conception of a monetary equilibrium as a tool for analysing economic development' (Lundberg 1937, p. 246; cf. Ohlin 1937, p. 224).

See Also

▶ Stockholm School

Bibliography

Hammarskjöld, D. 1933. Utkast till en algebraisk metod för dynamisk prisanalys. *Economisk* Tidskrift 34(5–6), 1932 (printed 1933), 157–176.

Hansen, B. 1951. *A study in the theory of inflation*. London: Allen & Unwin.

Hansson, B. 1981. *The Stockholm school and the development of dynamic method*. London: Croom Helm.

Hayek, F.A. 1931. *Prices and production*. London: Routledge & Sons. 2nd ed, 1935.

Hayek, F.A. 1933. Über 'neutrales' Geld. *Zeitschrift für Nationalökonomie* 4(5): 659–661. Quoted from and trans. as 'On "neutral" money', in F.A. Hayek, *Money, capital & fluctuations. Early essays*, ed. R. McCloughry, 159–162. London: Routledge & Kegan Paul, 1984.

Koopmans, F.G. 1933. Zum Problem des 'neutralen' Geldes. In *Beiträge zur Geldtheories*, ed. F.A. Hayek, 211–359. Vienna: Springer.

Lindahl, E. 1930. *Penningpolitikens medel*. Lund: Gleerup; enlarged version of 1st ed, 1929. Revised version trans. as Lindahl (1939a).

Lindahl, E. 1934. A note on the dynamic pricing problem. Mimeo, Gothenburg, 13 October. Quoted from the corrected version published in Steiger (1971), 204–211.

Lindahl, E. 1939a. The rate of interest and the price level. In *Studies in the theory of money and capital*, ed. E. Lindahl, 139–260. London: Allen & Unwin. Revised version of Lindahl (1930).

Lindahl, E. 1939b. Additional note (1939). Appendix to Lindahl (1939a), 260–268.

Lundberg, E. 1937. *Studies in the theory of economic expansion*. Stockholm: Norstedt & Söner.
Myrdal, G. 1932. Om penningteoretisk jämvikt. En studie över den 'normala räntan' i Wicksells penninglära. *Ekonomisk Tidskrift* 33(5–6), 1931 (printed 1932), 191–302. Revised version trans. as Myrdal (1933).
Myrdal, G. 1933. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie*, ed. F.A. Hayek, 361–487. Vienna: J. Springer. 1st revised version of Myrdal (1932); 2nd revised version trans. as Myrdal (1939).
Myrdal, G. 1939. *Monetary equilibrium*. London: Hodge. Revised version of Myrdal (1933).
Ohlin, B. 1937. Some notes on the Stockholm theory of savings and investment, II. *Economic Journal* 47: 221–240.
Palander, T. 1941. Om 'Stockholmsskolans' begrepp och metoder. Metodologiska reflexioner kring Myrdals 'Monetary Equilibrium'. *Ekonomisk Tidskrift* 43(1): 88–143. Quoted from and trans. as 'On the concepts and methods of the 'Stockholm School'. Some methodological reflections on Myrdal's 'Monetary Equilibrium', *International Economic Papers* No. 3, 1953, 5–57.
Shackle, G.L.S. 1945. Myrdal's analysis of monetary equilibrium. *Oxford Economic Papers* OS 7: 47–66.
Siven, C.-H. 1985. The end of the Stockholm school. *Scandinavian Journal of Economics* 87(4): 577–593.
Steiger, O. 1971. *Studien zur Entstehung der Neuen Wirtschaftslehre in Schweden. Eine Anti-Kritik*. Berlin: Duncker & Humblot.
Wicksell, K. 1936. *Geldzins und Güterpreise. Eine Studie über die den Tauschwert des Geldes bestimmenden Ursachen*. Jena: G. Fischer. Quoted from and trans. as Interest and prices. A study of the causes regulating the value of money. London: Macmillan, 1936.
Wicksell, K. 1903. Den dunkla punkten i penningteorien. *Ekonomisk Tidskrift* 5(12): 485–507.
Wicksell, K. 1935. *Föreläsningar i nationalekonomi. Vol. II: Om penningar och kredit*. Stockholmand Lund: Fritzes and Berlingska. Quoted from the trans. of the 3rd Swedish ed (1929), *Lectures on political economy*. Vol. II: *Money*. London: Routledge & Sons, 1935.

Monetary Overhang

Holger C. Wolf

Abstract

A monetary overhang emerges when individuals jointly hold more money than they wish and all adjustment processes are rendered unavailable through price and quantity controls. While monetary overhangs can in principle be eliminated through increased real money demand, their magnitude in practice typically implies a resolution through a reduction in real money supply through a cut in the nominal money supply or through higher prices. The former is impeded by the difficulty of estimating the appropriate reduction, the latter risks triggering sustained inflation in the presence of distorted relative wage and price structures.

Keywords

Forced saving; Inflation; Monetary overhang; Money supply; Price control; Price liberalization; Repressed inflation; Velocity of circulation

JEL Classifications

F3

In functioning market economies, an excess of nominal money supply over nominal money demand is resolved through a combination of price, interest rate and real income changes. If these adjustment mechanisms are effectively blocked, a *monetary overhang* may emerge. Periods of pervasive monetary overhangs occurred in 1940s Europe (Gurley 1953; Ames 1954; Dornbusch and Wolf 2001) and in the final period of some centrally planned economies, though for the latter episodes the magnitude of monetary overhangs – and thus the degree to which they contributed to rapid inflation in the aftermath of liberalization – has been debated (Nuti 1989; Cochrane and Ickes 1991; Chawluk and Cross 1997).

A pure monetary overhang requires three conditions. Individuals (*a*) face a binding upper limit on nominal expenditures on goods and services (typically reflecting rationing of goods at controlled prices), (*b*) face binding limits on the purchase of (non-monetary) assets, and (*c*) are holding monetary balances that exceed the levels they would choose to hold in the absence of restrictions on goods and asset purchases. In practice, for a number of reasons discussed below,

these constraints are unlikely to bind absolutely for all individuals; the term monetary overhang is hence also used more loosely to describe situations of extensive constraints on monetary spending.

First, access to unofficial markets may allow consumers a choice between converting monetary balances into goods at the higher unofficial price (hidden inflation) and holding cash balances (possibly in expectation of greater availability of rationed products at the lower official price in the future). As access to black markets is often limited and subject to penalties, the aggregate situation may still be described as a monetary overhang. Second, individuals may be able to convert cash into savings accounts. If interest rates are controlled, a situation may arise in which individuals prefer buying more goods at controlled prices to holding either cash or deposits, but, unable to buy goods, prefer the interest-bearing asset to cash. In this setting, the overhang situation persists, but now becomes a broader financial asset overhang (forced savings).

A monetary overhang – which might be alternatively characterized as a situation of excess nominal money supply, of below equilibrium prices (repressed inflation) and of below equilibrium velocity – can be eliminated by a combination of (*a*) a cut in the nominal money supply, (*b*) an increase in prices, (*c*) a decrease in equilibrium velocity, and (*d*) an increase in output.

In practice, the degree of disequilibrium is typically such that an increase in money demand through the third and fourth channel does not provide more than a partial solution. In episodes of often substantial uncertainty, higher nominal interest rates on demand deposits are unlikely to elicit pronounced increases in desired holdings and may, moreover, adversely affect stability in financial sectors often characterized by significant non-performing loans accrued during the period of price and interest rate controls. Rapid output growth following a return to free prices has at times acted as an anti-inflation force in a post-monetary-reform period, but rarely suffices to raise money demand sufficiently.

Severe monetary overhangs consequently tend to be cured by a reduction in the real money supply, either through an increase in the price level measured from the controlled price baseline (some black market prices may well fall after price liberalization) or through a cut in the nominal money supply (typically accompanied by the removal of price controls).

A cut in money supply (often embedded in a more comprehensive reform package) may be voluntary, for instance through the issue of bonds (with fiscal implications), or involuntary, either through a straight cancellation of part of the outstanding monetary balances or a forced conversion into public assets (again with associated fiscal implications). In principle, the cut in the nominal money supply can be set so that the post-reform equilibrium price level coincides with the pre-reform controlled price level. Determining the necessary cut requires estimates of the reform-induced change in velocity and output levels. The combination of extensive economic distortions in the pre-reform period, possible responses to anticipated monetary reform and the endogeneity of the post-reform developments to the success of the reform renders this estimation highly challenging. In economies with a recent market experience, historical velocity provides a useful baseline. Alternatively, velocity estimates can be based on comparable market economies. On the implementation side, the difficulty can be overcome by a two-stage approach combining an outright cancellation of part of the nominal money supply with a freeze on a further part, with an option to either cancel or release the frozen balances at a future point depending on the post-reform evolution of output and velocity.

Price liberalization relies on market forces to restore monetary equilibrium and avoids the need to estimate the extent of the overhang. If price controls kept prices for all goods below their equilibrium by the same proportion, the monetary overhang can in principle be resolved with a one-time proportionate jump in all prices. In practice, the disequilibrium price level typically combines with a disequilibrium relative price structure. Price liberalization may then lead to a period of inflation depending on the wage and price setting structures, possibly reinforced by an adverse fiscal impact of inflation.

See Also

- ▶ Command Economy
- ▶ Forced Saving
- ▶ Inflation
- ▶ Inflation Dynamics
- ▶ Rationing

Bibliography

Ames, E. 1954. Soviet bloc currency conversions. *American Economic Review* 44: 339–353.

Chawluk, A., and R. Cross. 1997. Measures of shortage and monetary overhang in the Polish economy. *Review of Economics and Statistics* 79: 105–115.

Cochrane, J., and B. Ickes. 1991. Inflation stabilization in reforming socialist economies. *Comparative Economic Studies* 33: 97–122.

Dornbusch, R., and H. Wolf. 2001. Curing a monetary overhang. In *Money, capital mobility, and trade: Essays in honor of Robert A. Mundell*, ed. G.A. Calvo, M. Obstfeld, and R. Dornbusch. Cambridge, MA: MIT Press.

Gurley, J. 1953. Excess liquidity and European monetary reforms. *American Economic Review* 43: 76–100.

Nuti, D. 1989. Hidden and repressed inflation in soviet type economies. *Contributions to Political Economy* 5: 37–82.

Monetary Policy

David E. Lindsey and Henry C. Wallich

The term *monetary policy* refers to actions taken by central banks to affect monetary and other financial conditions in pursuit of the broader objectives of sustainable growth of real output, high employment, and price stability. The average rate of growth of the stock of money in circulation has been viewed for centuries as the decisive determinant of overall price trends in the long run. General financial conditions associated with money creation or destruction, including changes in interest rates, also have been considered for some time an important factor of business cycles.

In the modern era, the bulk of money in developed economies consists of bank deposits rather than gold and silver or government-issued currency and coin. Accordingly, governments have authorized central banks today to guide monetary developments with instruments that afford control over deposit creation and affect general financial conditions. Central banks' actions are deliberately aimed at influencing the performance of the nation's economy and are not based on ordinary business considerations, such as profit. The guideposts and degree of discretion central banks should use in implementing monetary policy remain controversial issues, as are questions of the coordination of monetary policy with fiscal policy and with policies abroad.

The Instruments of Monetary Policy

The instruments available to central banks vary from country to country, depending on institutional structure, political system, and stage of development. In most developed capitalist economies, central banks basically use one or more of three main instruments to control deposit creation and affect financial conditions. *Required reserve ratios* set minimum fractions of certain deposit liabilities that commercial banks and in some countries thrift institutions must hold on reserve as assets in the form of cash in their vaults or deposits at the central bank. *The discount or official rate* is the interest charged by the central bank for providing reserve deposits directly to the banking system either through lending at a 'discount window' or through rediscounting or purchases of financial assets held by banks.

Open market operations are the third instrument. They involve either outright or temporary purchases and sales, typically of government securities, by the central bank with the market in general. The central bank pays for a securities purchase by crediting the reserve deposit account of the seller's bank, which in turn credits the deposit account of the seller. The central bank receives payment for a sale of securities by debiting the reserve account of the buyer's bank,

which in turn debits the account of the buyer. In this way, open market operations that alter the amount of securities held in the central bank's asset portfolio have as their counterpart a change in the nonborrowed reserves held by banks, that is, the reserves that do not originate through bank discount borrowings. The amount of these nonborrowed reserves also is changed by variations in other, noncontrolled items on the asset or liability side of the central bank's balance sheet, such as gold holdings that were important historically or the deposits of domestic and foreign governments that can vary considerably today. Still, central banks routinely monitor these items and can prevent them from having sizable undesired impacts on nonborrowed reserves by engaging in offsetting open market operations.

The sum of borrowed and nonborrowed reserves constitutes the total reserves available to the banking system. The central bank can exercise considerable control over these two sources of total reserve availability. Open market operations, as noted, provide for fairly close control of overall non-borrowed reserves. The level of the discount rate as well as other administrative procedures affect the amount of borrowed reserves. Given the interest rates on other sources of short-term bank funding, a change in the discount rate, or commonly in some countries in other lending terms and conditions, alters the incentives banks face to borrow reserves at the discount window. A discount rate increase, for example, would tend to induce banks to reduce their discount borrowing and turn to other sources of funds. Banks would attempt to replace the funds by borrowing reserves from other banks, or by issuing large-sized certificates of deposit, or even by selling liquid financial assets in secondary markets. These actions would transmit upward tendencies to the interest rates on these instruments.

The control by central banks over the availability of total reserves to private banks gives central banks at one remove a decisive influence over the availability of deposits to the public as well as over conditions in the money market. Given total reserves, the required reserve ratio sets an upper limit on the amount of deposits that can be created. In practice, this upper limit is not reached because private banks desire to hold a portion of total reserves not as required reserves but in the form of a cushion of reserves in excess of requirements. But since excess reserves are assets that typically earn no interest, unlike loans and investments, banks seek to hold them to minimal levels.

If reserves represent the lever central banks can use to control deposits, then the required reserve ratio represents the fulcrum. A given increase in the supply of total reserves has an amplified effect on deposits. This is the case whether it is brought about through an open market operation that tends to raise non-borrowed reserves or a cut in the discount rate that tends to raise borrowed reserves. Banks initially receiving the new reserves could immediately attempt to loan their surfeit of reserves to other banks, thus depressing the interest rate on overnight loans of reserves between banks. The easing of conditions in this market puts downward pressure on rates on other money market instruments, such as Treasury bills or large certificates of deposit. This general reduction in short-term interest rates encourages the public to hold more transactions and savings deposits, because the incentive to economize on such money balances is reduced by the narrower opportunity cost (in terms of foregone interest income) of holding low-return deposits instead of other interest-bearing assets. Deposits will rise, boosting required reserves, until required reserves have risen enough to exhaust all unwanted excess reserves, which necessitates an expansion in deposits that is some multiple of the original increase of reserves.

Required reserve ratios also represent a potentially active, alternate instrument for varying supplies of money and credit. Changes in these requirements alter the amount of bank deposits that a given quantity of total reserves can support. However, reserve requirement variations are a blunt instrument at best, as even relatively small changes in them produce large effects on the amount of deposits that can be supported by reserves outstanding. Accordingly, central banks infrequently resort to changes in these required reserve ratios.

Some countries do not impose reserve requirements. In those cases, the central bank's liabilities to banks are represented by voluntarily held vault cash and clearing or working balances. These central banks can still use open-market-type operations to influence deposit creation and money market conditions by varying reserve supply relative to these voluntary demands for reserves. However, the relationship between reserves and deposits, which in these countries depends on the average of the banks' desired ratios of reserve assets to deposits of the public, is less predictable than is the case with binding reserve requirements.

Whether the banking system's vault cash and deposits at the central bank are held predominantly as required or voluntary reserves, total reserves plus currency outside banks represent the nation's total monetary base. This aggregate also is potentially controllable by the central bank. Since currency has traditionally been supplied to meet the demands of the public, as a practical matter, however, central banks have found it more advisable to exercise direct control over reserves than over the monetary base.

Variations in the supply of reserves relative to the demand for them, with associated impacts on the cost of reserves, other interest rates, and the stock of money, are the initial channels through which most central banks of developed capitalist countries use their policy instruments to affect the macroeconomy. Some countries with less developed securities markets rely more heavily on policies focused on bank lending, including in some cases direct controls on bank credit through ceilings or reserve requirements against bank assets. The activities of these central banks in controlling aggregate credit and its allocation are conceptually separate from monetary policy *per se* and are not considered in this article.

The Distinctions Between Monetary Policy, Debt Management and Fiscal Policy

Monetary policy can be distinguished from debt management and fiscal policy. Debt management and monetary policy are similar only in the limited sense that both change the composition of the public's holdings of financial assets and the public's liquidity position through shifts between short- and longer-term assets. More liquidity is provided if the government shortens the average maturity of its debt outstanding. Similarly, if a central bank purchases government debt from a member of the public, liquidity is enhanced because the public has traded a less liquid security for a more liquid deposit. Nonetheless, an open market purchase by the central bank of government securities in effect retires the debt, by replacing securities outstanding in the hands of the banks or the public with bank reserves and associated public deposits, both of which earn no or below-market returns on the margin. The injection of this kind of reserve liability of the central bank from outside the private economy brings about widespread portfolio adjustments that lower market interest rates generally as an aspect of the expansion in money. A debt management operation of the federal government, by contrast, just replaces one security in the hands of the public with another, affecting the term structure of outstanding debt and possibly the term structure of interest rates but not the general level of interest rates.

Monetary policy is clearly distinguishable from fiscal policy because each affects the economy through a different route. Fiscal policy has a direct effect on spending through government outlays and a direct effect on income available for spending through tax rates. Fiscal policy also has a financial aspect because budgetary deficits or surpluses imply changes in government debt that presumably influence total credit demands and interest rates. (On the other hand, to the debatable extent that the public views government debt as entailing an ultimate tax liability, a larger government deficit indirectly would tend to encourage an equal and offsetting increase in private saving to finance future tax payments and hence discourage private spending.) In contrast to the direct spending and income effects of fiscal policy, the impact of monetary policy is wholly indirect and depends on the response of spenders and borrowers to the changes in monetary and financial conditions brought about by policy actions.

The Macroeconomic Effects and Objectives of Monetary Policy

Monetary policy responsibilities of central banks today go far beyond the role originally seen for central banks, which involved ensuring the stability of the banking system and the convertibility of deposits, especially in times of financial panics. Early in their history, central banks assumed the role of 'the lender of last resort', meaning that they would provide a source of funds for financially troubled banks to forestall liquidity crises. Subsequent experience indicated the need for central banks to provide an 'elastic currency' to accommodate seasonal variations in the demands for reserve assets. By doing so, central banks could avoid periodic reserve shortages that had disturbed market conditions and also, on occasion, confidence as well, giving rise to runs on banks. Deposit insurance, bank supervision with on-site examinations, and bank regulation ranging from circumscribing certain risky activities to setting minimum requirements for bank capital or certain bank assets or liabilities also have been introduced to help assure a stable banking system. In some countries, responsibility for many of these functions has been granted to other governmental agencies.

A major role for central banks in maintaining the safety and soundness of the financial system has continued to the present day, even though it has been joined in this century by a responsibility for overall macroeconomic stabilization. Macroeconomic stability requires a sound financial system; a weak financial system may not be able to withstand the effects of exogenous shocks to the economy or of restrictive policy actions that otherwise would be appropriate.

The dominant influence of monetary policy over time on the price level traditionally has elevated long-term price stability to a paramount position among the macroeconomic objectives of central banks. Under a gold standard historically, the world stock of gold provided a longer-term anchor to the world's average price level. But the commitment of central banks to buy and sell gold at a fixed price in terms of the domestic currency automatically gave rise to substantial inflows or outflows of gold to individual countries in the process of international adjustment. Large impacts on domestic economic activity and prices resulted in cyclical instability and sustained inflationary or deflationary episodes. The demise of the gold standard lessened the constraints on central banks in pursuing shorter-term domestic stabilization goals, but the discipline of the outstanding gold stock over long-term international price trends also was lost. In the modern era, central banks have been given the charge of exercising self-discipline in seeking the objective of longer-term price stability. Meanwhile, the widely recognized short-run impact of monetary policy on economic activity and employment has fostered increased emphasis on countercyclical objectives as well.

Over extended periods, the effects of monetary policy are concentrated almost wholly on nominal magnitudes, that is, those measured in terms of the monetary unit. As noted, central banks are able to control the nominal stock of bank reserves and, at one remove, the money stock. Average price trends become established as the nominal quantity of money interacts over time with the private sector's demand for real money balances, that is, the value of money after adjustment for the impact of inflation or deflation of prices. Thus, monetary policy has considerable influence over the long run on the average price level. In addition, factors that affect demands for real money balances, such as financial innovation, and more generally, that affect demands or supplies of aggregate output also play a role in price level determination.

The supply of output is determined in the long run mainly by real factors such as population growth, participation in the labour force, capital accumulation, and productivity trends. Real values for wages, interest rates, and currency exchange rates also respond secularly to fundamental real forces. The influence of monetary policy over the level and trend rate of change of the nominal price level carries over indirectly as an influence on the nominal values of these other variables but not on their real values. The real values of wages, interest rates, and exchange rates that are ground out by the market economy interact over time with nominal price behaviour to determine their nominal values. In the very long run, then, a change in the nominal quantity of money will be neutral as all nominal

prices and wages tend to adjust proportionally, *ceteris paribus*.

While the influence of monetary policy on the behaviour of real values is widely agreed to be minor over the long pull, it is also recognized that monetary policy can affect real variables significantly in a shorter run, cyclical context. Doubts about the effectiveness of expansive monetary policy under conditions of a domestic depression raised during the Keynesian revolution have since been largely resolved. The views of today's mainstream macroeconomists with regard to the impact of monetary impulses on real economic activity are not far from those expressed in the following passage from David Hume:

> Though the high prices of commodities be a necessary consequence of the encrease in gold and silver, yet it follows not immediately upon that encrease; but some time is required before the money circulates through the whole state and makes its effect be felt on all ranks of people. At first no alteration is perceived; by degrees the price rises, first of one commodity then another; till the whole at last reaches a just proportion with the new quantity of specie . . ..In my opinion, it is only this interval, or intermediate situation, between the acquisition of money and rise of prices, that the encreasing quantity of gold and silver is favourable to industry (David Hume, 'Of Money', 1752; reprinted in *Writings on Economics*, edited by Eugene Rotwein, Madison: University of Wisconsin Press, 1955).

The proposition that monetary policy actions necessarily have a short-run effect on real variables is not universally accepted. In the last decade, the macro rational expectations school has argued that changes in monetary policy may not alter real variables, even in the short run. If a policyinduced movement in the nominal money stock is expected by the public in advance, then the public will have the incentive to adjust accordingly the actual, as well as expected, levels of all nominal values. Such a public response in principle would neutralize even the short-run impact of the expected policy change on real variables.

This recent challenge to the traditional view concedes, though, that unexpected policy actions can alter real variables, if only temporarily. Unanticipated policy actions can cause the outcomes for various nominal, and thus real, magnitudes to diverge, at least for a time, from their expected values. But the rational expectations school stresses that the public will come to expect policy actions that respond systematically to economic developments. Only policy actions that were purely random, or based on information not shared by the public, would then be unexpected, in which case the scope for effective countercyclical policy would be greatly narrowed.

In recent years, however, considerable counterevidence has been marshalled to the view that only unexpected policy moves can affect real values. Most empirical studies suggest that even systematic and expected changes in the direction of monetary policy do not show through fully right away in nominal values but have short-run impacts on real economic values.

The evident lagged effects on nominal values have been explained by various frictions, adjustment costs, and information imperfections. While prices may adjust minute-by-minute in auction markets, in other markets explicit or implicit longer-term contracts impart rigidities to nominal prices and wages, preventing a complete short-run adjustment to even expected changes in nominal policy variables. Costs of changing certain prices also can give rise to gradual adjustment of nominal magnitudes over time. In addition, the buffer role of inventories keeps even an expected change in nominal spending on goods and services from being felt by all producers simultaneously. Finally, because firms and workers get information about demands for their own goods and services more rapidly than information about economy-wide demands, they can misperceive as only local events what really are generalized phenomena ultimately affecting all nominal values. Economic agents can be induced in the short-run to change their real behaviour in supplying goods and services, rather than fully altering the nominal prices or wages they offer as would actually be called for by overall developments.

The Channels Through which Monetary Policy Affects the Economy

Even though economists now better understand these general behaviour patterns, the precise

channels through which monetary policy actions are transmitted to the economy at large and the specific variables that best indicate the stance of monetary policy remain unresolved issues. The immediate effects of changes in the instruments controlled by central banks on the supply and cost of reserves are clear. Both an open market purchase of government securities that raises nonborrowed reserves and a cut in the discount rate augment reserve availability relative to demands for excess and required reserves. This places interest rates on money market instruments under downward pressure. After that, an almost infinite sequence of 'ripple effects' ensues, and analysts still differ in sorting out the most important of these in affecting the economy. Their differing views reflect the complexity of the linkages between the modern financial system and economic activity and the alternative simplifications various schools have adopted in an effort to capture the essential elements.

The mainstream view derives from the Keynesian tradition and highlights induced movements in market interest rates across the maturity spectrum as the primary linkage between monetary policy actions and private spending. An 'easing' or 'tightening' of monetary policy is indexed by decreases or increases in market rates. Of course, the distinction between nominal and real interest rates is recognized; a change in market interest rates that simply compensates for an accompanying change in inflationary expectations may have minimal real economic effects.

These Keynesian channels of influence have been worked out in some detail, both theoretically and in large-scale econometric models. With an easing monetary policy action, for example, the initial fall in money market rates induces market participants to revise downward their expected levels of future short-term rates as well, causing a softening in long-term rates. Inflation expectations are thought to adjust sluggishly in lagged response to actual inflation and to be largely unresponsive to the monetary easing itself. Thus, any tendency for inflation expectations to rise and mute the decline in nominal longer-term rates is viewed as minor. More administered interest rates, such as the prime rate and consumer credit and mortgage rates also come under downward pressure over time, and credit terms and conditions tend to become less restrictive.

Spending in the interest-sensitive sectors, such as housing, consumer durables, and business investment, are most affected at first, as lowered borrowing costs stimulate demand. Some second-round effects also begin to come into play. The associated increase in income and production further stimulates consumption and investment spending. Also, the fall in interest rates is mirrored by a rise in financial asset values, and this gain in wealth encourages even more consumption spending.

Prices come under delayed upward pressure in part because tighter labour markets reduce the unemployment rate, at least transitionally, below its 'natural' level consistent with the realization of wage and price expectations. Such a fall in unemployment is associated with an acceleration of wage rates. Higher capacity utilization also may boost price markups over costs. As the actual inflation rate picks up, inflation expectations begin to increase as well, imparting a separate upward thrust to price and wage setting.

An internationally related channel also can become important, especially in countries with a significant external sector and flexible exchange rates. A more accommodative monetary policy action that reduces domestic interest rates is likely to diminish the demand for assets denominated in the home currency. Under flexible exchange rates, the resulting depreciation of the exchange value of the currency will lower export prices in world markets and raise import prices. These developments will work over time to bolster spending on net exports. But as the associated rise in import prices feeds through the domestic price structure, broad price indexes also will tend to move higher.

Monetarists adopt a somewhat different viewpoint, asserting that monetary policy stimulus is best measured by the growth of the money stock. A sustained speed-up in money growth after some lag leads to a temporary strengthening in real economic activity and even later to faster inflation. The process is set in motion as an injection of reserves supports more money than the public desires to hold given prevailing levels of real

income, prices and interest rates. As the extra balances 'burn a hole in people's pockets', purchases of a wide variety of goods and services as well as financial assets are stimulated. Short-term market interest rates may fall initially, but more importantly, prices across a broad spectrum of financial and real assets are bid up, stimulating demand for and production of investment and consumer goods. Monetarists, like Keynesians, contend that in the long run the impact on real activity dissipates as the monetary stimulus becomes fully reflected in inflation. People end up needing the extra money just to carry out normal transactions at inflated prices, leaving no more extra stimulus to real spending.

Guides for Monetary Policy

With a wide variety of financial and non-financial measures affected in the process of economic adjustment to a monetary policy action, the question remains as to which variable represents the best indicator of the stance of policy, that is, the variable providing the most reliable indication of the future effects of monetary policy on the economy. Moreover, with policy decisions having lagged effects and policymakers necessarily uncertain about economic linkages and trends, such a variable presumably also could be used to keep policymakers' judgement from going astray by serving as an intermediate guide to monetary policy actions. An intermediate guide is a variable that the central bank would attempt to keep in line with a prespecified target, and thus it would need to be reasonably controllable by the central bank. The central bank would adjust the level of the intermediate target less frequently than the settings of the policy instruments.

Central banks over time have used, with evolving emphasis, alternative primary policy guides. Historically, the price at which gold or some other metal was convertable into the domestic currency played this role. Subsequently, market interest rates and foreign exchange rates received more emphasis as policy guides. In recent decades, targets for overall money and debt have been adopted in many industrial countries. Other candidates have been proposed, including the monetary base, indexes of commodity prices or the general price level, nominal GNP, and real interest rates.

Unfortunately, both macroeconomic analysis and experience suggest that no single variable can consistently serve as a reliable policy guide, so no hard-and-fast answer as to the best one can be given that holds under all conditions. All variables beyond non-borrowed reserves and the discount rate are influenced by factors other than monetary policy actions, and it turns out that the degree of stimulus to the economy involved in movements in any of them will depend on the nature of the other factors at work. Summarizing the advantages and disadvantages of several variables demonstrates this dilemma.

Monetary aggregates represent collections of financial assets, grouped according to their degree of 'moneyness'. Narrow measures of money comprise currency and fully checkable deposits to encompass the public's primary transactions balances. Broader measures also include other highly liquid accounts with additional savings features. Sharp lines of demarcation separating the various aggregates are difficult to draw as the characteristics of various assets often shade into one another over a wide spectrum, especially in countries with developed, deregulated, and innovative financial markets.

Monetary aggregates serve well as policy guides when the public's demands for them are stably related to nominal spending and market interest rates and have a relatively small interest sensitivity. Suppose, for example, that there is a cyclical downturn in total spending. If the central bank withdraws reserves from the system in order to maintain a given level of market interest rates in the face of falling demand for money, the money stock would decrease at a time when additional monetary stimulus is needed. If instead the central bank maintains the original level of reserves in order to keep the money stock at its target level, interest rates must fall. The less interest-sensitive is money demand, the more would interest rates have to decline to offset the depressing effect of reduced spending on the public's desired money holdings. Thus, by maintaining money at the

target level, an easing of credit conditions and perhaps a depreciating foreign exchange rate over time would partially offset the original decline in spending, and moderate the cyclical downturn.

However, when the public's willingness to hold monetary aggregates given nominal spending and interest rates is undergoing an abnormal shift, movements in measures of the money stock provide misleading signals of monetary stimulus or restraint. Such shifts in money demand have occurred in response to financial innovations and deposit deregulation as well as varying precautionary motives on the part of the public. As a result, the properties of empirical relationships connecting the money stock to nominal spending and market interest rates have been altered – in some cases permanently. The precise nature of the impact is difficult to assess when the process is underway. For example, in the United States during the 1980s, the disinflation process interacted with sluggishly adjusting offering rates on newly deregulated transaction deposits to raise substantially the responsiveness of the demand for narrow money to changes in market interest rates. The sizeable declines in market interest rates after the early 1980s enhanced the relative attractiveness of returns on interest-bearing fully checkable deposits, which are included in narrow money. Inflows into these accounts were massive, with a significant portion representing savings-type funds.

Faced with unusual money demand behaviour, the central bank would be best advised not to resist departures of money from target but instead to accommodate reserve provision to the shifting demands for money. It could do so by maintaining existing reserve market conditions. Otherwise, the very process of restoring the money stock to target would transmit the disturbance in money demand to spending behaviour and economic activity. The changing conditions in reserve and credit markets associated with returning money to target would be inappropriate for stabilizing spending. Central banks that rely on monetary aggregates as policy guides have interpreted such episodes as demonstrating the need for monitoring overall economic developments and making feedback adjustments to monetary targets in response to evident disturbances of money demand relative to income.

Market interest rates thus would serve as a better policy guide than monetary aggregates if the only disturbances were to the money demand relationship. In a realistic economic context, though, independent disturbances to the relation between nominal spending and market interest rates also are likely to occur. Collection lags for data on economic activity and uncertainties about the structure of behavioural relations in the economy and the permanence of disturbances make the appropriate reaction to unexpected pressures on interest rates and misses of money from target difficult to determine at the time. For example, suppose the central bank sees that an unanticipated rise in interest rates is needed to keep the money stock from overshooting its target. The reason could be an unexpected strengthening of inflation and nominal spending that is boosting money demand, or a surprise upward shift in money demand relative to spending, or some combination of the two. The source of overshoot of money from target could prove self-reversing, or it could be only the beginning of a cumulative departure. Unless uncertainty about the money demand relationship is exceptionally severe, it might be safer for the central bank to permit some upward movement in nominal interest rates than for it simply to keep interest rates stable by fully accommodating reserve provision to the outsized money growth. The latter reaction would provide no counterweight at all to what later could prove to have been an inflationary upturn of nominal spending.

On the other hand, suppose spending had clearly weakened, and the central bank has responded by adding to reserve availability in the face of a very interest sensitive demand for the targeted monetary aggregate. The resulting fall in interest rates has led to a sizeable overshoot of money from target. In this circumstance, it may turn out better for the economy if the central bank accepts the full overrun of money above target. With a highly interest-sensitive demand for money, only a small reduction in interest rates is implied by keeping money on target when spending turns down. This easing in financial conditions

will provide only little offset to the weakness in economic activity, unless there is an upward adjustment to targeted money growth.

Relying more on interest rates as a policy guide will not necessarily resolve the problem of determining the appropriate central bank reaction to unexpected developments. The relationship between nominal values of spending and market interest rates is qualitatively less predictable and stable over time than the already loose underlying relation between their real values. Determining what level of real interest rates is associated with a given level of nominal interest rates is hampered because the public's inflationary expectations are difficult to measure. Longerterm real interest rates, which are thought to have the most powerful influence on many important components of real spending, are especially difficult to discern since the public's expectations of inflation over the distant future are the most obscure.

Central banks thus face considerable uncertainty about the real interest rate that would be implied initially by the choice of a particular level for the nominal interest rate. Also, unless the resulting level of real interest rates just happened to be consistent with full employment and a stable inflation rate, the implied real interest rate would tend to move over time in a destabilizing direction, as was originally pointed out by Knut Wicksell. Suppose the central bank maintained nominal interest rates over an extended period at a level that from the start yielded an overly stimulative real interest rate. Economic activity would press against the economy's productive and labour capacities, and inflation would tend to accelerate. But as inflation expectations adjusted upward in response to actual inflation, the real interest rate implied by the targeted nominal interest rate would be driven still lower. This fall in the real interest rate would add even more stimulus to nominal spending and inflation. Even so, growth of reserves and money would have to be continually accelerated to maintain the targeted nominal interest rate. An ever faster rise in nominal spending and inflation hence would result from pegging the nominal interest rate at too low a level. Those central banks emphasizing market interest rates as policy guides interpret such possibilities as requiring them to monitor overall monetary and economic developments and to make feedback adjustments over time in setting market interest rates.

Since the potential pitfalls of either monetary aggregates or market interest rates as policy guides have induced central banks to respond to more ultimate gauges of economic performance – such as nominal GNP, prices, and unemployment – in setting intermediate targets, some observers have recommended that central banks should cut through the feedback process by simply targeting one of these ultimate objectives itself. But this approach has disadvantages beyond the fact that the particular objective variable to be selected is of course controversial. Any of these ultimate variables are affected by numerous forces outside the central bank's control, including domestic fiscal policy and foreign fiscal and monetary policies. Data on most of these variables are received with some delay and then subject to sizeable revisions. Finally, an attempt to convert an ultimate objective to a shorter-term policy target would risk unstable macroeconomic outcomes over time in light of the uncertainties and lags in the impact of money policy actions.

For these reasons, central banks have not believed that they can justifiably be held accountable for the near-term performance of the overall economy. Despite the problems of interpreting the various monetary and debt aggregates and interest rates which are more under their near-term control, central banks, as well as many other analysts, view the constellation of these financial variables taken together as offering a surer indication of the longer-term stance of monetary policy itself than current values of ultimate economic variables. While the disadvantages under some circumstances of guiding policy by any single financial measure argue against an overreliance on any one, the advantages of each under different circumstances are viewed as suggesting that, when taken in the context of broader economic developments as well, none can be completely ignored in the conduct, or assessment, of monetary policy.

Nevertheless, the long-run linkage between money growth and inflation together with the traditional concern of central banks for price stability give monetary aggregates a special position

among these financial variables. Continuing to focus on average money growth over extended periods, while accounting for the influence of distortions to its demand behaviour, forces central banks to keep longer-term price objectives in mind in the process of adjusting policy actions in response to shorter-term financial and economic developments.

Policy Rules Versus Discretion

Some critics of the discretion embodied in such a policy approach place even more weight on the longer-term consideration of providing a nominal anchor to the macroeconomy. They also interpret the difficulty of forecasting both economic developments and the impact of policy actions as implying that central banks should not even attempt to stabilize the economy over shorter periods of time through discretionary policy actions. Given the lags and uncertainties involved, they believe such flexibility in policy is likely to do more harm than good, despite the best of intentions.

These critics have recommended that monetary policy should be based on fixed rules rather than discretion. The most influential has been the proposal of the monetarists to maintain a low, constant money growth rate through thick and thin. These economists, under the intellectual leadership of Milton Friedman, have argued that excessive money growth is the main cause of inflation and that variations in monetary growth historically have been responsible for the large cyclical fluctuations in real output. With constant money growth, self-correcting mechanisms would prevent macroeconomic shocks from having major, sustained impacts on economic activity.

The rational expectations school has added a new wrinkle to the case for policy rules. They believe discretion imparts an inflationary bias to monetary policy because central banks face an irresistible temptation over time to put aside announced long-term plans to maintain price stability in pursuit of short-term production and employment aims. If the public had adjusted price expectations to the central bank's announced intention to maintain price stability, then a temporary increase in money growth would surprise the public and cause a desirable, if short-lived boost to output and employment with little inflationary cost. But with rational expectations, the public would see through this temptation and expect such a policy action. Expectations of inflation would emerge in anticipation of the monetary stimulus, leaving only price increases but no output gains as the policy is implemented. Indeed, if the central bank did not undertake the expected stimulus after all, then output would instead be temporarily depressed. Given this dilemma, central banks would end up providing the monetary stimulus, even though it only validates ongoing inflation and has no output effects.

Following an invariant policy rule would avoid this problem, according to these advocates, by making an anti-inflation policy credible to the public. The public then would expect only policy actions consistent with price stability. This school supports a rule defined in terms of a fixed target for either money growth or the price level.

While monetarist views have affected central bank practice in recent decades, as evidenced by the enhanced reliance on monetary aggregates in actual policy making during the 1970s, central banks have shied away from the adoption of fixed money rules in light of the perceived advantages of policy flexibility. The abstract, even hypothetical, nature of the rational expectations argument has limited its influence. And the substantial disinflation worldwide from the early- to mid-1980s despite continued rapid growth of monetary aggregates appears to have weakened the case of both schools for policy rules.

Coordination with Other Domestic and Foreign Policies

The separate influences of domestic fiscal policy and foreign fiscal and monetary policies on macroeconomic outcomes at home raise the issue of coordination with domestic monetary policy. On the domestic side, a more expansionary fiscal policy involving enlarged government spending or reduced taxes, for example, may require that

offsetting actions be taken to make monetary policy more restrictive. Even if the policy mix is changed in such a way to keep overall employment, production and prices the same, nominal and real values of market interest rates and foreign exchange rates would be altered, as would the composition of aggregate output in terms of real consumption, investment and net exports.

The traditional view has been that after some point a shift in the policy mix toward more stimulative fiscal policy and more restrictive monetary policy becomes undesirable, since investment and net exports will have to be 'crowded out' by higher real interest rates and exchange rates to make room for larger government purchases or private consumption. A reduced pace of investment would retard capital accumulation and the economy's longer-term growth potential, while lowered net exports would harm export and import-competing industries. The increased government budget deficit would be associated with a larger deficit in the current international payments accounts, implying a faster buildup of both government and external debt. Repayments of both debts over time would become more burdensome for domestic residents by requiring a greater sacrifice of future consumption. If capital inflows were invested effectively, they could provide resources to make future debt-service payments, but if these funds simply helped to finance government budget deficits, they would not support private capital accumulation.

A more recently advanced 'supply side' view is that sizeable reductions in marginal tax rates will encourage private saving, investment, work effort and entrepreneurship. The economy's growth potential will be increased sufficiently that a more restrictive monetary policy need not be adopted, even if government deficits initially are increased. Evidence drawn from the United States following sizeable cuts in marginal tax rates early in the 1980s suggests, however, that the resulting incentive effects on the economy's potential growth rate are relatively minor.

In practice, fiscal policy has not proven to be as flexible a macroeconomic tool as monetary policy, as other social goals beyond countercyclical considerations, as well as legislative delays, have prevented prompt adjustment in spending programmes or tax laws in response to overall economic developments. This situation has placed monetary policy in the forefront in pursuing macroeconomic stabilization objectives. Monetary policy actions become most politically sensitive when fiscal policy is expansionary and private spending and wage and price decisions are causing the economy to overheat. The required turn to a more restrictive monetary policy engenders opposition to higher interest rates, particularly from sectors where employment and production are especially disadvantaged by upward movements in interest and exchange rates. Having monetary policy bear too much of the brunt of countercyclical policy restraint is to be avoided partly because the central bank may not practically be able to bear the political pressures, and partly because economic imbalances across sectors become more pronounced.

Difficulties of achieving the proper mix of monetary and fiscal policy are exacerbated when considered in a multi-country context. International policy coordination is not just an issue of meshing monetary policies, but of coordinating overall macro-policy mixes in general. It also covers a range of possible interactions among countries. A higher degree of policy coordination obviously becomes more necessary in a regime of fixed exchange rates or common trade areas, or to the degree that different countries have accepted common exchange rate objectives. But even without explicit exchange rate objectives, some international policy coordination may still yield benefits given the transmission of effects of policy actions. A general move to restrictive fiscal policy abroad, for example, would reduce foreign spending on domestic exports. Also, the fall in foreign interest rates can heighten the willingness of international investors to hold domestic financial assets; these higher asset demands would act to keep domestic interest rates lower than otherwise but raise the exchange rate, ultimately depressing further the domestic balance of trade. Self-reinforcing cycles can even occur in which more expansive fiscal policies abroad, with a rise in foreign interest rates, produce a depreciation of the exchange value of the domestic currency. The

lower value of the currency then leads to higher domestic inflation and inflationary expectations, in turn possibly contributing to a further depreciation of the currency, depending on the domestic monetary policy response.

A process of international policy coordination is in the interest of interrelated nations. Closer coordination could in principle provide for a greater measure of stability in exchange markets, while maintaining some of the features of flexible exchange rates in cushioning international disturbances and in lessening the constraints on policy implied by automatic flows of international reserves under a fixed-rate system. But the interests and circumstances of sovereign nations may well diverge at times. This can occur either because of a somewhat different emphasis on the various ultimate economic objectives or because the countries are experiencing different stages of the business cycle. In such situations, scope for agreement about the appropriate pattern of macroeconomic policies across countries may be limited.

See Also

- ▶ Budgetary Policy
- ▶ Central Banking
- ▶ International Monetary Policy
- ▶ Monetarism

Bibliography

Axilrod, S.H. 1985. U.S. monetary policy in recent years: An overview. *Federal Reserve Bulletin* 71(1): 14–24.

Bank of England. 1984. *The development and operation of monetary policy, 1960–1983*. London: Oxford University Press.

Friedman, M. 1960. *A program for monetary stability*. New York: Fordham University Press.

Goodheart, C.A.E. 1984. *Monetary theory and practice: The UK experience*. London: Macmillan.

Lindsey, D.E. 1986. The monetary regime of the Federal Reserve System. In *Alternative monetary regimes*, ed. C.D. Campbell and W.R. Dougen. Baltimore: Johns Hopkins University Press.

McCallum, B.T. 1984. Credibility and monetary policy. In *Price stability and public policy*. Kansas City: Federal Reserve Bank of Kansas City.

Poole, W. 1970. Optimal choice of monetary policy instruments in a simple stochastic macro model. *Quarterly Journal of Economics* 84(2): 197–216.

Wallich, H.C., and P.M. Keir. 1979. The role of operating guides in U.S. monetary policy: A historical review. *Federal Reserve Bulletin* 65(9): 679–691.

Monetary Policy, History of

Michael D. Bordo

Abstract

Monetary policy has evolved over the centuries, with the development of the money economy. To implement monetary policy the monetary authority uses its policy instruments (short-term interest rates or the monetary base) to achieve its goals of low inflation and real output close to potential. This article surveys the origins of monetary policy from the classical gold standard to the evolution of central banks and their quest for goal independence, documenting the evolution of the goals, instruments and intermediate targets of monetary policy, and surveying the development of theories of monetary policy, including the debate over rules versus discretion.

Keywords

Assignats; Bank of England; Bank rate; Banking school; Bretton Woods system; Bullionist controversy; Central bank independence; Commitment; Convertibility; Credit rationing; Currency school; Debasement; Discount rate; Federal reserve system; Fiat money; Fiduciary money; Fiscal policy; Gold standard; Great depression; Inflation; Inflation targeting; Interest rate; Law, J; Lender of last resort; Monetarism; Monetary base central bank; Monetary policy targets; Monetary policy, history of; Money; Money supply; Natural rate of interest; Open market operations; Phillips curve; Price stability; Quantity theory of money; Real bills doctrine; Real rate of return on capital;

Rediscounting; Reserve requirements; Rules versus discretion; Seigniorage; Sterilization; Taylor rule; Time consistency; Velocity of circulation

JEL Classification
D4; D10

Today monetary policy is the principle way in which governments influence the macroeconomy. To implement monetary policy the monetary authority uses its policy instruments (short-term interest rates or the monetary base) to achieve its desired goals of low inflation and real output close to potential. Monetary policy has evolved over the centuries, along with the development of the money economy.

The Origins

Debate swirls between historians, economists, anthropologists and numismatists over the origins of money. In the West it is commonly believed that coins first appeared in ancient Lydia in the eighth century BC. Some date the origins to ancient China.

Money evolved as a medium of exchange, a store of value and unit of account. According to one authority – Hicks (1969), following Menger (1892) – its rise was associated with the growth of commerce. Traders would hold stocks of another good, in addition to the goods they traded in, which was easily stored, widely recognized, and divisible, with precious metals evolving as the best example. This good would serve as the unit of account and then as a medium of exchange. According to this story money first emerged from market activity.

Governments became involved when the monarch realized that it was easier to pay his soldiers in terms of generalized purchasing power than with particular goods. This led to the origin of seigniorage or the government's prerogative in the coining of money. Seigniorage originally represented the fee that the royal mint collected from the public to convert their holdings of bullion into coin. Governments generally since ancient times had a monopoly over the issue of coins (either licensing their production or producing them themselves).

The earliest predecessors to monetary policy seem to be those of debasement, where the government would call in the coins, melt them down and mix them with cheaper metals. They would alter either the weight or the quality of the coins (fineness). An alternative method used was to alter the unit of account (see Redish 2000; Sussman 1993; Sargent and Velde 2002). The practice of debasement was widespread in the later years of the Roman Empire (Schwartz 1973), but reached its perfection in western Europe in the late Middle Ages. Sussman (1993) describes how the French monarchs of the fifteenth century, unable to collect more normal forms of taxes, used debasement as a form of inflation tax to finance the ongoing Hundred Years War with the English. Debasement was really a form of fiscal rather than monetary policy, but it set the stage for the later development of monetary policy using fiduciary money.

Fiduciary or paper money evolved from the operations of early commercial banks in Italy (Cipolla 1967) to economize on the precious metals used in coins (although there is evidence that paper money was issued by imperial decree in China centuries earlier: see Chown 1994). This development has its origins in the practice of goldsmiths who would issue warehouse receipts as evidence of their storing gold coins and bullion for their clients. Eventually these certificates circulated as media of exchange. Once the goldsmiths learned that not all the claims were redeemed at the same time, they were able to circulate claims of value greater than their specie reserves. Thus was borne fiduciary money (money not fully backed by specie) and fractional reserve banking. The goldsmiths and early commercial bankers learned by experience to hold a precautionary reserve sufficient to meet the demands for redemption in the normal course of business.

Governments began issuing paper money in Europe only in the eighteenth century. An early example was Sweden's note issue, initiated to finance its participation in the Seven Years War

(Eagly 1969). Fiat money reached its maturity during the American Revolutionary Wars when the Congress issued continentals to finance military expenditures. These were promissory notes to be convertible into specie; but the promise was not kept. They were issued in massive quantities. However, the rate of issue and the average inflation rate of 65% per annum (Rockoff 1984) was not far removed from the revenue-maximizing rate of issue by a monopoly fiat money issuing central bank of the twentieth century (Bailey 1956). During the French Revolution the overissue of paper money, the *assignats*, which were based initially on the value of seized Church lands, led to hyperinflation (White 1995).

An early predecessor of monetary policy was John Law's system. In 1719 Law persuaded the Regent of France to convert the French national debt into stock in his Compagnie des Indes. He then used the stock as backing for the issue of notes in his Banque Royale. Note issue could then support and finance the issue of further shares. Law then conducted a proto typical form of monetary policy in 1720 to save his system when he attempted both to peg the exchange rate of notes in terms of specie and provide a support price to stem the collapse in the price of shares (Bordo 1987; Velde 2007).

Central Banks

Monetary policy is conducted by the monetary authority. It is the issuer of national currency and the source of the monetary base. Usually we think of central banks as fulfilling these functions, but in many countries, until well into the twentieth century, in the absence of a central bank, these were performed by the Treasury or in some cases (Australia, Canada, New Zealand) by a large commercial bank entrusted with the government's tax revenues (Goodhart 1989). The earliest central banks were established in the seventeenth century (the Swedish Riksbank founded in 1664, the Bank of England founded in 1694, the Banque de France, founded in 1800, and the Netherlands Bank in 1814) to aid the fisc of the newly emerging nation states.

In the case of the Bank of England a group of private investors was granted a royal charter to set up a bank to purchase and help market government debt. The establishment of the bank helped ensure the creation of a deep and liquid government debt market which served as the base of growing financial system (Dickson 1969; Rousseau and Sylla 2003. The bank eventually evolved into a bankers' bank by taking deposits from other nascent commercial banks. Its large gold reserves and monopoly privilege eventually allowed it to become a lender of last resort, that is, to provide liquidity to its correspondents in the face of a banking panic – a scramble by the public for liquidity.

Monetary policy as we know it today began by the bank discounting the paper of other financial institutions, both government debt and commercial paper. The interest rate at which the bank would lend, based on this collateral became known as bank rate (in other countries as the discount rate). By altering this rate the bank could influence credit conditions in the British economy. It could also influence credit conditions in the rest of the world by attracting or repelling short-term funds (Sayers 1957).

A second wave of central banks was initiated at the end of the nineteenth century. This was not based explicitly on the fiscal revenue motive as had been the case with the first wave, but on following the rules of the gold standard and ironing out swings in interest rates induced by seasonal forces and by the business cycle. Included in this group are the Swiss National Bank founded in 1907 (Bordo and James 2007) and the Federal Reserve founded in 1913 (Meltzer 2003). Subsequent waves of new central banks followed in the interwar period as countries in the British Empire, the new states of central Europe and Latin America attempted to emulate the experiences of the advanced countries (Capie et al. 1994).

Central Bank Independence

Although the early central banks had public charters, they were privately owned and they had policy independence. A problem that plagued

the Bank of England in its early years was that it placed primary weight on its commercial activities and on several occasions of financial distress was criticized for neglecting the public good. Walter Bagehot formulated the responsibility doctrine in 1873 according to which the bank was to place primary importance on its public role as lender of last resort (Bagehot 1873).

From the First World War onwards central banks focused entirely on public objectives, and many fell under public control. Their objectives also changed from emphasis on maintaining specie convertibility towards shielding the domestic economy from external shocks and stabilizing real output and prices. This trend continued in the 1930s and after the Second World War. Moreover, the Great Depression led to a major reaction against central banks, which were accused of creating and exacerbating the depression. In virtually every country monetary policy was placed under the control of the Treasury and fiscal policy became dominant. In every country central banks followed a low interest peg to both stimulate the economy and aid the Treasury in marketing its debt.

Monetary policy was restored to the central banks in the 1950s (for example, in the United States, after the Treasury–Federal Reserve Accord of 1951), and there followed a brief period of price stability until the mid-1960s. This was followed by a significant run up in inflation worldwide. The inflation was broken in the early 1980s by concerted tight monetary policies in the United States, the United Kingdom and other countries and a new emphasis placed on the importance of low inflation based on credible monetary policies. Central banks in many countries were granted goal independence and were given a mandate to keep inflation low.

Classical Monetary Policy

The true origin of modern monetary policy occurred under the classical gold standard, which prevailed from 1880 to 1914. The gold standard evolved from the earlier bimetallic regime. Under the gold standard all countries would define their currencies in terms of a fixed weight of gold and then all fiduciary money would be convertible into gold. The key role of central banks was to maintain gold convertibility. Central banks were also supposed to use their discount rates to speed up the adjustment to external shocks to the balance of payments, that is, they were supposed to follow the 'rules of the game' (Keynes 1930). In the case of a balance of payments deficit, gold would tend to flow abroad and reduce a central bank's gold reserves. According to the rules, the central bank would raise its discount rate. This would serve to depress aggregate demand and offset the deficit. At the same time the rise in rates would stimulate a capital inflow. The opposite set of policies was to be followed in the case of a surplus.There is considerable debate on whether the rules were actually followed (Bordo and MacDonald 2005). There is evidence that central banks sterilized gold flows and prevented the adjustment mechanism from working (Bloomfield 1959). Others paid attention to the domestic objectives of price stability or stable interest rates or stabilizing output (Goodfriend 1988). There is also evidence that because the major central banks were credibly committed to maintaining gold convertibility they had some policy independence to let their interest rates depart from interest rate parity and to pursue domestic objectives (Bordo and MacDonald 2005).

After the First World War the gold standard was restored, but in the face of a changing political economy – the extension of suffrage and organized labour (Eichengreen 1992) – greater emphasis was placed by central banks on the domestic objectives of price stability and stable output and employment than on external convertibility. Thus for example the newly created Federal Reserve sterilized gold flows and followed countercyclical policies to offset two recessions in the 1920s (Meltzer 2003).

The depression beginning in 1929 was probably caused by inappropriate monetary policy. The Federal Reserve followed the flawed real bills doctrine, which exacerbated the downturn, and the gold sterilization policies followed by the Fed and the Banque de France greatly weakened the adjustment mechanism of the gold standard.

As mentioned above, the central banks were blamed for the depression and monetary policy was downgraded until the mid-1950s.

The Goals of Monetary Policy

The goals of monetary policy have changed across monetary regimes. Until 1914, the dominant monetary regime was the gold standard. Since then the world has gradually shifted to a fiat money regime. Under the classical gold standard the key goal was gold convertibility with limited focus on the domestic economy. By the interwar period gold convertibility was being overshadowed by emphasis on domestic price level and output stability, and the regime shifted towards fiat money. This continued after the Second World War. Under the 1944 Bretton Woods Articles of Agreement, member countries were to maintain pegged exchange rates and central banks were to intervene in the foreign exchange market to do this, but the goal of domestic full employment was also given predominance. The Bretton Woods system evolved into a dollar gold exchange standard in which member currencies were convertible on a current account basis into dollars and the dollar was convertible into gold (Bordo 1993). A continued conflict between the dictates of internal and external balance was a dominant theme from 1959 to 1971 as was the concern over global imbalance because the United States, as centre country of the system, would provide through its balance of payments deficits and its role as a financial intermediary more dollars than could be safely backed by its gold reserves (Triffin 1960).

The collapse of Bretton Woods between 1971 and 1973 was brought about largely because the United States followed an inflationary policy to finance both the Vietnam War and expanded social welfare programmes like Medicare under President Johnson's Great Society, thus ending any connection of the monetary regime to gold and propelling the world to a pure fiat regime. In this new environment the balance was largely tipped in favour of domestic stability and was coupled with the now dominant belief by central bankers in the Phillips curve trade-off between unemployment and inflation (Phillips 1958): this led to a focus on maintaining full employment at the expense of inflation.

The resulting 'great inflation' of the 1970s finally came to an end in the early 1980s by central banks following tight monetary policies. Since then the pendulum has again swung towards the goal of low inflation and the belief that central banks should eschew control of real variables (Friedman 1968; Phelps 1968).

The Instruments of Monetary Policy

The original policy instrument was the use of the discount rate and rediscounting. Open market operations (the buying and selling of government securities) was first developed in the 1870s and 1880s by the Bank of England in order to make bank rate effective, that is to force financial institutions to borrow (Sayers 1957). Other countries with less developed money markets than those of Britain used credit rationing (France) and gold policy operations to alter the gold points and impede the normal flow of gold (Sayers 1936).

In the interwar period the newly established Federal Reserve initially used the discount rate as its principal tool, but after heavy criticism for its use in rolling back the post-First World War inflation and thereby creating one of the worst recessions of the twentieth century in 1920–1921 (Meltzer 2003), the Fed shifted to open market policy, its principal tool ever since. In the 1930s it also began changing reserve requirements. Its policy of doubling reserve requirements in 1936 was later blamed as the cause for the recession of 1937–1938 (Friedman and Schwartz 1963). In the 1930s and 1940s, along with the downgrading of monetary policy, came an increased use of various types of controls and regulations such as margin requirements on stock purchases, selective credit controls on consumer durables and interest rate ceilings. Similar policies were adopted elsewhere. The return to traditional monetary policy in the 1950s restored open market operations to the position of predominance.

Intermediate Targets

Traditionally, central banks altered interest rates as the mechanism to influence aggregate spending, prices and output. In the 1950s, the monetarists revived the quantity theory of money and posited the case for using money supply as the intermediate target (Friedman 1956; Brunner and Meltzer 1993). The case for money was based on evidence of a stable relationship between the growth of money supply, on the one hand, and nominal income and the price level, on the other hand, and the evidence that, by focusing on interest rates, the Fed and other central banks aggravated the business cycle, and then – in part because of their inability to distinguish between real and nominal rates – generated the great inflation of the 1970s (Brunner and Meltzer 1993).

By the 1970s most central banks had monetary aggregate targets. However, the rise in inflation in the 1970s (which was followed by disinflation) as well as continuous financial innovation (which was in turn exacerbated by inflation uncertainty) made the demand for money function less predictable (Laidler 1980; Judd and Scadding 1982). This meant that central banks had difficulty in meeting their money growth targets. In addition, the issue was raised as to which monetary aggregate to target (Goodhart 1984). By the late 1980s most countries had abandoned monetary aggregates and returned to interest rates. But since the early 1990s monetary policy in many countries has been based on pursuing an inflation target (implicit or explicit) with the policy rate set to allow inflation to hit the target, a policy which seems to be successful.

Theories of Monetary Policy

The development of the practice of monetary policy described above was embedded in major advances in monetary theory that began in the first quarter of the nineteenth century. A major controversy in England, the Currency Banking School debate, has shaped subsequent thinking on monetary policy ever since. That debate evolved out of the Bullionist debate during the Napoleonic wars over whether inflation in Britain was caused by monetary or real forces (Viner 1937). In a later debate, Currency School advocates emphasized the importance for the Bank of England to change its monetary liabilities in accordance with changes in its gold reserves – that is, according to the currency principle, which advocated a rule tying money supply to the balance of payments. The opposing Banking School emphasized the importance of disturbances in the domestic economy and the domestic financial system as the key variables the Bank of England should react to. They advocated that the bank directors should use their discretion rather than being constrained by a rigid rule. The controversy still rages.

Later in the nineteenth century, the two principles became embedded in central banking lore (Meltzer 2003, ch. 2). The Federal Reserve and other central banks (including the Swiss National Bank) were founded on two pillars that evolved from this debate – the gold standard and the real bills doctrine.

The latter evolved from nineteenth-century practice and the Banking School theory. The basic premise of real bills is that as long as commercial banks lend on the basis of self-liquidating short-term real bills they will be sound. Moreover, as long as central banks discount only eligible real bills the economy will always have the correct amount of money and credit. Adherence to real bills sometimes clashed with the first pillar, gold adherence, for example when the economy was expanding and real bills dictated ease while the balance of payments was deteriorating, which dictated tightening. This conflict erupted in the United States on a number of occasions in the 1920s (Friedman and Schwartz 1963).

Adherence to the two pillars led to disaster in the 1930s. The Fed made a serious policy error by following real bills. A corollary of that theory urged the Fed to defuse the stock market boom because it was believed that speculation would lead to inflation, which would ultimately lead to deflation (Meltzer 2003). According to Friedman and Schwartz, Meltzer and others, the Fed's tight policy triggered a recession in 1929 and its inability to stem the banking panics that followed in the

early 1930s led to the Great Depression. The depression was spread globally by the fixed exchange rate gold standard. In addition, the gold standard served as 'golden fetters' for most countries because, lacking the credibility they had before 1914, they could not use monetary policy to allay banking panics or stimulate the economy lest it trigger a speculative attack (Eichengreen 1992).

The Great Depression gave rise to the Keynesian view that monetary policy was impotent. This led to the dominance of fiscal policy over monetary policy for the next two decades. The return to traditional monetary policy in the 1950s was influenced by Keynesian monetary theory. According to this approach monetary policy should influence short-term rates and then by a substitution process across the financial portfolio would affect the real rate of return on capital. This money market approach dominated policy until the 1960s.

The monetarists criticized the Fed for failing to stabilize the business cycle, for still adhering to vestiges of real bills (for example, free reserves: Calomiris and Wheelock 1998), and for its belief in a stable Phillips curve – that unemployment could be permanently reduced at the expense of inflation. This, they argued, led to an acceleration of inflation as market agents' expectations adjusted to the higher inflation rate, which produced the great inflation of the 1970s. As mentioned above, the subsequent adoption of monetary aggregate targeting was short lived because of unpredictable shifts in velocity.

The approach to monetary policy followed since the early 1990s has learned the basic lesson from the monetarists of the primacy of price stability. It also learned about the distinction between nominal and real interest rates (Fisher 1922). Moreover, it has adopted a principle from the earlier gold standard literature, Wicksell's (1898) distinction between the natural rate of interest and the bank rate (Woodford 2003). In Wicksell's theory, central banks should gear their lending rate to the natural rate (the real rate of return on capital). If it keeps bank rate too low, inflation will ensue, which under the gold standard will lead to gold outflows and upward market pressure on the bank rate. Today's central banks, dedicated to low inflation, can be viewed as following the Taylor rule, according to which they set the nominal policy interest rate relative to the natural interest rate as a function of the deviation of inflation forecasts from their targets and real output from its potential (Taylor 1999).

Rules Versus Discretion

A key theme in the monetary policy debate is the issue of rules versus discretion. The question that followed the Currency Banking School debate was whether monetary policy should be entrusted to well meaning authorities with limited knowledge or to a rule that cannot be designed to deal with unknown shocks (Simons 1936; Friedman 1960).

A more recent approach focuses on the role of time inconsistency. According to this approach a rule is a credible commitment mechanism that ties the hands of policymakers and prevents them from following time-inconsistent policies – policies that take past policy commitments as given and react to the present circumstances by changing policy (Kydland and Prescott 1977; Barro and Gordon 1983). In this vein, today's central bankers place great emphasis on accountability and transparency to support the credibility of their commitments to maintain interest rates geared towards low inflation (Svensson 1999).

Conclusion

Monetary policy has evolved since the early nineteenth century. It played a relatively minor role before 1914, although it was then that many of its tools and principles were developed. The role of monetary policy in stabilizing prices and output came to fruition in the 1920s, but for the Federal Reserve, which used a flawed model – the real bills doctrine – and adhered to a less than credible gold standard, the policy was a recipe for disaster and led to the great contraction of 1929–1933. When monetary policy was restored in the 1950s in the United States, it still was influenced by real

bills (Calomiris and Wheelock 1998), which may have led to the policy mistakes that created the great inflation. The rest of the world was tied to the United States by the pegged exchange rates of Bretton Woods. Since the early 1990s monetary policy in many countries has returned back to a key principle of the gold standard era – price stability based on a credible nominal anchor (Bordo and Schwartz 1999) and to Wicksell's distinction between real and nominal interest rates. Yet it is based on a fiat regime and the commitment of central banks to follow credible and predictable policies.

See Also

- ▶ Bank of England
- ▶ Central Bank Independence
- ▶ Fiat Money
- ▶ Gold Standard
- ▶ Inflation Targeting
- ▶ Monetary and Fiscal Policy Overview

Bibliography

Barro, R.I., and D.B. Gordon. 1983. Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics* 12: 101–121.

Bloomfield, A.I. 1959. *Monetary policy under the international gold standard*. New York: Federal Reserve Bank of New York.

Bagehot, W. 1873. *Lombard Street: A description of the money market*. Reprint edn. London: John Murray, 1917.

Bailey, M.J. 1956. The welfare costs of inflationary finance. *Journal of Political Economy* 64: 93–110.

Bordo, M.D. 1987. John Law. In *The new Palgrave: A dictionary of economic theory and doctrine*, ed. J. Eatwell and M. Milgate. London: Macmillan.

Bordo, M.D. 1993. The Bretton Woods international monetary system: A historical overview. In *A Retrospective on the Bretton Woods system: Lessons for international monetary reform*, ed. M.D. Bordo and B. Eichengreen. Chicago: University of Chicago Press.

Bordo, M.D. and James, H. 2007. The SNB 1907–1946: A happy childhood or a troubled adolescence? In *Swiss National Bank*. Centenary Conference volume, Zurich.

Bordo, M.D., and R. MacDonald. 2005. Interest rate interactions in the classical gold standard: 1880–1914: Was there monetary independence? *Journal of Monetary Economics* 52: 307–327.

Bordo, M.D., and A.J. Schwartz. 1999. Monetary policy regimes and economic performance: The historical record. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woolford. New York: North-Holland.

Brunner, K., and A.H. Meltzer. 1993. *Money and the economy: Issues in monetary analysis*. Cambridge, UK: Cambridge University Press.

Calomiris, C.W., and D.C. Wheelock. 1998. Was the great depression a watershed for American monetary policy? In *The defining moment: The great depression and the American economy in the twentieth century*, ed. M.D. Bordo, C. Goldin, and E.N. White. Chicago: University of Chicago Press.

Capie, F., C. Goodhart, S. Fischer, and N. Schnadt. 1994. *The future of central banking*. Cambridge, UK: Cambridge University Press.

Chown, J.F. 1994. *The history of money from AD 800*. London: Routledge.

Cipolla, C.M. 1967. *Money, prices, and civilization in the Mediterranean world, fifth to seventeenth century*. New York: Gordian Press.

Dickson, P.M. 1969. *The financial revolution in England: A study in the development of public credit, 1688–1756*. London: Macmillan.

Eagly, R.U. 1969. Monetary policy and politics in mid-eighteenth century Sweden. *Journal of Economic History* 29: 739–757.

Eichengreen, B. 1992. *Golden fetters*. New York: Oxford University Press.

Fisher, I. 1922. *The purchasing power of money*, 1965. New York: Augustus M. Kelley.

Friedman, M. 1956. Quantity theory of money: A restatement. In *Studies in the quality theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Friedman, M. 1960. *A program for monetary stability*. New York: Fordham University Press.

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.

Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Goodfriend, M. 1988. Central banking under the gold standard. *Carnegie Rochester Conference Series on Public Policy* 19: 85–124.

Goodhart, C.A.E. 1984. Chapter 3: Problems of monetary management. In *Monetary theory and practice. The UK experience*. London: Macmillan.

Goodhart, C. 1989. *The evolution of central banks*. Cambridge, MA: MIT Press.

Hicks, J.R. 1969. *A theory of economic history*. Oxford: Clarendon.

Judd, J.P., and J.L. Scadding. 1982. The search for a stable money demand function: A survey of the post-1973 literature. *Journal of Economic Literature* 20: 993–1023.

Keynes, J.M. 1930. *A treatise on money*, vol. 2: *The applied theory of money*. Repr. in *The collected writings of John Maynard Keynes*. 30 vols, ed. A. Robinson

and D. Moggridge, vol. 6. London: Macmillan for the Royal Economic Society, 1971.
Kydland, F.E., and E.C. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–492.
Laidler, D. 1980. The demand for money in the United States – Yet again. In *The state of macro-economics, Carnegie-Rochester conference series on public policy*, vol. 12, ed. K. Brunner and A.H. Meltzer. New York: North-Holland.
Meltzer, A.H. 2003. *A history of the federal reserve*, vol. 1. Chicago: University of Chicago Press.
Menger, K. 1892. On the origins of money. *Economic Journal* 2: 238–258.
Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom 1861–1957. *Economica* 25: 283–299.
Phelps, E.S. 1968. Money-wage dynamics and labor market equilibrium. *Journal of Political Economy* 76: 678–711.
Redish, A. 2000. *Bimetallism: An economic and historical analysis*. Cambridge, UK: Cambridge University Press.
Ricardo, D. 1811. High price of bullion: A proof of the depreciation of bank notes. In *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge, UK: Cambridge University Press.
Rockoff, H. 1984. *Drastic measures: A history of wage and price controls in the United States*. New York: Cambridge University Press.
Rousseau, P., and R. Sylla. 2003. Financial systems, economic growth and globalization. In *Globalization in historical perspective*, ed. M.D. Bordo, A. Taylor, and J. Williamson. Chicago: University of Chicago Press.
Sargent, T., and F. Velde. 2002. *The big problem of small change*. Princeton: Princeton University Press.
Sayers, R.S. 1936. *Bank of England operations, 1890–1914*. London: P.S. King & Son.
Sayers, R.S. 1957. *Central banking after Bagehot*. Oxford: Oxford University Press.
Schwartz, A.J. 1973. Secular price change in historical perspective. *Journal of Money, Credit, and Banking* 5: 243–269.
Simons, H.C. 1936. Rule versus authorities in monetary policy. *Journal of Political Economy* 44: 1–30.
Sussman, N. 1993. Debasement, royal reviews and inflation in France during the second stage of the Hundred Years War. *Journal of Economic History* 56: 789–808.
Svensson, L.E.O. 1999. Inflation targeting as a monetary policy rule. *Journal of Monetary Economics* 43: 607–654.
Taylor, J.B. 1999. A historical analysis of monetary policy rules. In *Monetary policy rule*, ed. J.B. Taylor. Chicago: University of Chicago Press.
Thornton, H. 1802. *An inquiry into the national effects of the paper credit of Great Britain*. Fairfield: Augustus M. Kelley, 1978.
Triffin, R. 1960. *Gold and the dollar crisis*. New Haven: Yale University Press.
Velde, F. 2007. *Government equity and money: John Laws system in 1720 France*. Princeton: Princeton University Press.
Viner, J. 1937. *Studies in the theory of international trade*. New York: Augustus M. Kelley, 1975.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy.* Princeton: Princeton University Press.
Wicksell, K. 1898. *Interest and prices*. New York: Augustus M. Kelley, 1965.
White, E.N. 1995. The French Revolution and the politics of government finance, 1770–1815. *Journal of Economic History* 55: 227–255.

Monetary Transmission Mechanism

Peter N. Ireland

Abstract

The monetary transmission mechanism describes how policy-induced changes in the nominal money stock or the short-term nominal interest rate impact on real variables such as aggregate output and employment. Specific channels of monetary transmission operate through the effects that monetary policy has on interest rates, exchange rates, equity and real estate prices, bank lending, and firm balance sheets. Recent research on the transmission mechanism seeks to understand how these channels work in the context of dynamic, stochastic, general equilibrium models.

Keywords

Asset price channels of monetary transmission; Balance sheet channel of monetary transmission; Balance sheet credit channel; Bank lending credit channel; Bank reserves; Bonds; Central banks; Currency; Exchange rate channel of monetary transmission; Inflation targeting; Interest rate channel of monetary transmission; Interest rates; IS–LM model; Keynesianism; Life-cycle theory of consumption; Liquidity trap; Loanable funds theory; Monetarism; Monetary base; Monetary policy; Monetary transmission mechanism; New

Keynesian economics; New Keynesian Phillips curve; Open market operations; Phillips curve; Rational expectations models; Real business cycles; Taylor rule

JEL Classifications
E4

The monetary transmission mechanism describes how policy-induced changes in the nominal money stock or the short-term nominal interest rate impact on real variables such as aggregate output and employment.

Key Assumptions

Central bank liabilities include both components of the monetary base: currency and bank reserves. Hence, the central bank controls the monetary base. Indeed, monetary policy actions typically begin when the central bank changes the monetary base through an open market operation, purchasing other securities – most frequently, government bonds – to increase the monetary base or selling securities to decrease the monetary base.

If these policy-induced movements in the monetary base are to have any impact beyond their immediate effects on the central bank's balance sheet, other agents must lack the ability to offset them exactly by changing the quantity or composition of their own liabilities. Thus, any theory or model of the monetary transmission mechanism must assume that there exist no privately issued securities that substitute perfectly for the components of the monetary base. This assumption holds if, for instance, legal restrictions prevent private agents from issuing liabilities having one or more characteristics of currency and bank reserves.

Both currency and bank reserves are nominally denominated, their quantities measured in terms of the economy's unit of account. Hence, if policy-induced movements in the nominal monetary base are to have real effects, nominal prices must not be able to respond immediately to those movements in a way that leaves the real value of the monetary base unchanged. Thus, any theory or model of the monetary transmission mechanism must also assume that some friction in the economy works to prevent nominal prices from adjusting immediately and proportionally to at least some changes in the monetary base.

The Monetary Base and the Short-Term Nominal Interest Rate

If, as in the US economy today, neither component of the monetary base pays interest or if, more generally, the components of the monetary base pay interest at a rate that is below the market rate on other highly liquid assets such as short-term government bonds, then private agents' demand for real base money M/P can be described as a decreasing function of the short-term nominal interest rate i: $M/P = \mathrm{L}(i)$. This function L summarizes how, as the nominal interest rate rises, other highly liquid assets become more attractive as short-term stores of value, providing stronger incentives for households and firms to economize on their holdings of currency and banks to economize on their holdings of reserves. Thus, when the price level P cannot adjust fully in the short run, the central bank's monopolistic control over the nominal quantity of base money M also allows it to influence the short-term nominal interest rate i, with a policy-induced increase in M leading to whatever decline in i is necessary to make private agents willing to hold the additional volume of real base money and, conversely, a policy-induced decrease in M leading to a rise in i. In the simplest model where changes in M represent the only source of uncertainty, the deterministic relationship that links M and i implies that monetary policy actions can be described equivalently in terms of their effects on either the monetary base or the short-term nominal interest rate.

Poole's (1970) analysis shows, however, that the economy's response to random shocks of other kinds can depend importantly on whether the central bank operates by setting the nominal quantity of base money and then allowing the market to determine the short-term nominal interest rate or by setting the short-term nominal interest

rate and then supplying whatever quantity of nominal base money is demanded at that interest rate. More specifically, Poole's analysis reveals that central bank policy insulates output and prices from the effects of large and unpredictable disturbances to the money demand relationship by setting a target for i rather than M. Perhaps reflecting the widespread belief that money demand shocks are large and unpredictable, most central banks around the world today – including the Federal Reserve in the United States – choose to conduct monetary policy with reference to a target for the short-term nominal interest rate as opposed to any measure of the money supply. Hence, in practice, monetary policy actions are almost always described in terms of their impact on a short-term nominal interest rate – such as the federal funds rate in the United States – even though, strictly speaking, those actions still begin with open market operations that change the monetary base.

The Channels of Monetary Transmission

Mishkin (1995) usefully describes the various channels through which monetary policy actions, as summarized by changes in either the nominal money stock or the short-term nominal interest rate, impact on real variables such as aggregate output and employment.

According to the traditional Keynesian *interest rate channel*, a policy-induced increase in the short-term nominal interest rate leads first to an increase in longer term nominal interest rates, as investors act to arbitrage away differences in risk-adjusted expected returns on debt instruments of various maturities as described by the expectations hypothesis of the term structure. When nominal prices are slow to adjust, these movements in nominal interest rates translate into movements in real interest rates as well. Firms, finding that their real cost of borrowing over all horizons has increased, cut back on their investment expenditures. Likewise, households facing higher real borrowing costs scale back on their purchases of homes, automobiles and other durable goods. Aggregate output and employment fall. This interest rate channel lies at the heart of the traditional Keynesian textbook IS–LM model, due originally to Hicks (1937), and also appears in the more recent New Keynesian models described below.

In open economies, additional real effects of a policy-induced increase in the short-term interest rate come about through the *exchange rate channel*. When the domestic nominal interest rate rises above its foreign counterpart, equilibrium in the foreign exchange market requires that the domestic currency gradually depreciate at a rate that, again, serves to equate the risk-adjusted returns on various debt instruments, in this case debt instruments denominated in each of the two currencies– this is the condition of uncovered interest parity. Both in traditional Keynesian models that build on Fleming (1962); Mundell (1963) and Dornbusch (1976) and in the New Keynesian models described below, this expected future depreciation requires an initial appreciation of the domestic currency that, when prices are slow to adjust, makes domestically produced goods more expensive than foreign-produced goods. Net exports fall; domestic output and employment fall as well.

Additional *asset price channels* are highlighted by Tobin's (1969) q-theory of investment and Ando and Modigliani's (1963) life-cycle theory of consumption. Tobin's q measures the ratio of the stock market value of a firm to the replacement cost of the physical capital that is owned by that firm. All else equal, a policy-induced increase in the short-term nominal interest rate makes debt instruments more attractive than equities in the eyes of investors; hence, following a monetary tightening, equilibrium across securities markets must be re-established in part through a fall in equity prices. Facing a lower value of q, each firm must issue more new shares of stock in order to finance any new investment project; in this sense, investment becomes more costly for the firm. In the aggregate across all firms, therefore, investment projects that were only marginally profitable before the monetary tightening go unfunded after the fall in q, leading output and employment to decline as well. Meanwhile, Ando and Modigliani's life-cycle theory of consumption assigns a role to wealth as well as income as key determinants of consumer

spending. Hence, this theory also identifies a channel of monetary transmission: if stock prices fall after a monetary tightening, household financial wealth declines, leading to a fall in consumption, output and employment.

According to Meltzer (1995), asset price movements beyond those reflected in interest rates alone also play a central role in *monetarist* descriptions of the transmission mechanism. Indeed, monetarist critiques of the traditional Keynesian model often start by questioning the view that the full thrust of monetary policy actions is completely summarized by movements in the short-term nominal interest rate. Monetarists argue instead that monetary policy actions impact on prices simultaneously across a wide variety of markets for financial assets and durable goods, but especially in the markets for equities and real estate, and that those asset price movements are all capable of generating important wealth effects that impact, through spending, on output and employment.

Two distinct *credit channels*, the *bank lending channel* and the *balance sheet channel*, also allow the effects of monetary policy actions to propagate through the real economy. Kashyap and Stein (1994) trace the origins of thought on the bank lending channel back to Roosa (1951) and also highlight Blinder and Stiglitz's (1983) resurrection of the loanable funds theory and Bernanke and Blinder's (1988) extension of the IS–LM model as two approaches that account for this additional source of monetary non-neutrality. According to this lending view, banks play a special role in the economy not just by issuing liabilities – bank deposits – that contribute to the broad monetary aggregates but also by holding assets – bank loans – for which few close substitutes exist. More specifically, theories and models of the bank lending channel emphasize that for many banks, particularly small banks, deposits represent the principal source of funds for lending and that for many firms, particularly small firms, bank loans represent the principal source of funds for investment. Hence, an open market operation that leads first to a contraction in the supply of bank reserves and then to a contraction in bank deposits requires banks that are especially dependent on deposits to cut back on their lending, and firms that are especially dependent on bank loans to cut back on their investment spending. Financial market imperfections confronting individual banks and firms thereby contribute, in the aggregate, to the decline in output and employment that follows a monetary tightening.

Bernanke and Gertler (1995) describe a broader credit channel, the balance sheet channel, where financial market imperfections also play a key role. Bernanke and Gertler emphasize that, in the presence of financial market imperfections, a firm's cost of credit, whether from banks or any other external source, rises when the strength of its balance sheet deteriorates. A direct effect of monetary policy on the firm's balance sheet comes about when an increase in interest rates works to increase the payments that the firm must make to service its floating rate debt. An indirect effect arises, too, when the same increase in interest rates works to reduce the capitalized value of the firm's long-lived assets. Hence, a policy-induced increase in the short-term interest rate not only acts immediately to depress spending through the traditional interest rate channel, it also acts, possibly with a lag, to raise each firm's cost of capital through the balance sheet channel, deepening and extending the initial decline in output and employment.

Recent Developments

Recent theoretical work on the monetary transmission mechanism seeks to understand how the traditional Keynesian interest rate channel operates within the context of dynamic, stochastic, general equilibrium models. This recent work builds on early attempts by Fischer (1977) and Phelps and Taylor (1977) to combine the key assumption of nominal price or wage rigidity with the assumption that all agents have rational expectations so as to overturn the policy ineffectiveness result that McCallum (1979) associates with Lucas (1972) and Sargent and Wallace (1975). This recent work builds on those earlier studies by deriving the key behavioural equations of the New Keynesian model from more detailed descriptions of the objectives and constraints faced by optimizing households and firms.

More specifically, the basic New Keynesian model consists of three equations involving three variables: output y_t, inflation π_t, and the short-term nominal interest rate i_t. The first equation, which Kerr and King (1996) and McCallum and Nelson (1999) dub the expectational IS curve, links output today to its expected future value and to the *ex ante* real interest rate, computed in the usual way by subtracting the expected rate of inflation from the nominal interest rate:

$$y_t = E_t y_{t+1} - \sigma(i_t - E_t \pi_{t+1}),$$

where σ, like all of the other parameters to be introduced below, is strictly positive. This equation corresponds to a log-linearized version of the Euler equation linking an optimizing household's intertemporal marginal rate of substitution to the inflation-adjusted return on bonds, that is, to the real interest rate. The second equation, the New Keynesian Phillips curve, takes the form

$$\pi_t = \beta E_t \pi_{t+1} + \gamma y_t$$

and corresponds to a log-linearized version of the first-order condition describing the optimal behavior of monopolistically competitive firms that either face explicit costs of nominal price adjustment, as suggested by Rotemberg (1982), or set their nominal prices in randomly staggered fashion, as suggested by Calvo (1983). The third and final equation is an interest rate rule for monetary policy of the type proposed by Taylor (1993),

$$i_t = \alpha \pi_t + \psi y_t,$$

according to which the central bank systematically adjusts the short-term nominal interest in response to movements in inflation and output. This description of monetary policy in terms of interest rates reflects the observation, noted above, that most central banks today conduct monetary policy using targets for the interest rate as opposed to any of the monetary aggregates. A money demand equation could be appended to this three-equation model, but that additional equation would serve only to determine the amount of money that the central bank and the banking system would need to supply to clear markets, given the setting for the central bank's interest rate target (see Ireland 2004, for a detailed discussion of this last point).

In this benchmark New Keynesian model, monetary policy operates through the traditional Keynesian interest rate channel. A monetary tightening in the form of a shock to the Taylor rule that increases the short-term nominal interest rate translates into an increase in the real interest rate as well when nominal prices move sluggishly due to costly or staggered price setting. This rise in the real interest rate then causes households to cut back on their spending, as summarized by the IS curve. Finally, through the Phillips curve, the decline in output puts downward pressure on inflation, which adjusts only gradually after the shock.

Importantly, however, the expectational terms that enter into the IS and Phillips curves displayed above imply that policy actions will differ in their quantitative effects depending on whether these actions are anticipated or unanticipated; hence, this New Keynesian model follows the earlier rational expectations models of Lucas and Sargent and Wallace by stressing the role of expectations in the monetary transmission mechanism. And, as emphasized by Kimball (1995), by deriving these expectational forms for the IS and Phillips curves from completely spelled-out descriptions of the optimizing behaviour of households and firms, the New Keynesian model takes advantage of the powerful microeconomic foundations introduced into macroeconomics through Kydland and Prescott's (1982) real business cycle model while also drawing on insights from earlier work in New Keynesian economics as exemplified, for instance, by the articles collected in Mankiw and Romer's (1991) two-volume set.

Clarida et al. (1999) and Woodford (2003) trace out the New Keynesian model's policy implications in much greater detail. Obstfeld and Rogoff (1995) develop an open-economy extension in which the exchange rate channel operates together with the interest rate channel of monetary transmission. Andres et al. (2004) enrich the New Keynesian specification to open up a broader

range of asset price channels and, similarly, Bernanke et al. (1999) extend the basic model to account for the balance sheet channel of monetary transmission. Hence, all of these papers contribute to a large and still growing body of literature that examines the workings of various channels of monetary transmission within dynamic, stochastic, general equilibrium models.

Other recent research on the monetary transmission mechanism focuses on the problem of the zero lower bound on nominal interest rates – a problem that appears most starkly in the basic New Keynesian model sketched out above, in which monetary policy affects the economy exclusively through the Keynesian interest rate channel. Private agents always have the option of using currency as a store of value; hence, equilibrium in the bond market requires a non-negative nominal interest rate. In a low-inflation environment where nominal interest rates are also low on average, the central bank may bump up against this zero lower bound and find itself unable to provide further monetary stimulus after the economy is hit by a series of adverse shocks. Interest in the zero lower bound grew during the late 1990s and early 2000s when, in fact, nominal interest rates approached zero in Japan, the United States and a number of other countries. Among recent studies, Summers (1991); Fuhrer and Madigan (1997) rank among the first to call for renewed attention to the problem of the zero lower bound; Krugman (1998) draws parallels between the zero lower bound and the traditional Keynesian liquidity trap; and Eggertsson and Woodford (2003); Svensson (2003), and Bernanke et al. (2004) propose and evaluate alternative monetary policy strategies for coping with the zero lower bound.

Finally, on the empirical front, quite a bit of recent work looks for evidence of quantitatively important credit channels of monetary transmission. Kashyap and Stein (1994); Bernanke et al. (1996) survey this branch of the literature. Also, the striking rise in equity and real estate prices that began in the mid-1990s in the United States, the United Kingdom, and elsewhere has sparked renewed interest in quantifying the importance of the asset price channels described above. Noteworthy contributions along these lines include Lettau and Ludvigson (2004); Case et al. (2005).

See Also

- ▶ Inflation Dynamics
- ▶ Liquidity Trap
- ▶ Monetary Business Cycles (Imperfect Information)
- ▶ Monetary Business Cycle Models (Sticky Prices and Wages)
- ▶ Money Supply
- ▶ Phillips Curve (New Views)
- ▶ Taylor Rules

Acknowledgment I would like to thank Steven Durlauf and Jeffrey Fuhrer for extremely helpful comments and suggestions.

Bibliography

Ando, A., and F. Modigliani. 1963. The 'life cycle' hypothesis of saving: Aggregate implications and tests. *American Economic Review* 53: 55–84.

Andres, J., J. Lopez-Salido, and E. Nelson. 2004. Tobin's imperfect asset substitution in optimizing general equilibrium. *Journal of Money, Credit, and Banking* 36: 665–690.

Bernanke, B., and A. Blinder. 1988. Credit, money, and aggregate demand. *American Economic Review* 78: 435–439.

Bernanke, B., and M. Gertler. 1995. Inside the black box: The credit channel of monetary policy transmission. *Journal of Economic Perspectives* 9(4): 27–48.

Bernanke, B., M. Gertler, and S. Gilchrist. 1996. The financial accelerator and the flight to quality. *The Review of Economics and Statistics* 78: 1–15.

Bernanke, B., M. Gertler, and S. Gilchrist. 1999. The financial accelerator in a quantitative business cycle framework. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Bernanke, B., V. Reinhart, and B. Sack. 2004. Monetary policy alternatives at the zero bound: An empirical assessment. *Brookings Papers on Economic Activity* 2004(2): 1–78.

Blinder, A., and J. Stiglitz. 1983. Money, credit constraints, and economic activity. *American Economic Review* 73: 297–302.

Calvo, G. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

Case, K., J. Quigley, and R. Shiller. 2005. Comparing wealth effects: The stock market versus the housing

market. *Advances in Macroeconomics* 5: 1. http://www.bepress.com/bejm/advances/vol5/iss1/art1/.
Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy: A New Keynesian perspective. *Journal of Economic Literature* 37: 1661–1707.
Dornbusch, R. 1976. Expectations and exchange rate dynamics. *Journal of Political Economy* 84: 1161–1176.
Eggertsson, G., and M. Woodford. 2003. The zero bound on interest rates and optimal monetary policy. *Brookings Papers on Economic Activity* 2003(1): 139–211.
Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.
Fleming, J. 1962. Domestic financial polices under fixed and under floating exchange rates. *International Monetary Fund Staff Papers* 9: 369–379.
Fuhrer, J., and B. Madigan. 1997. Monetary policy when interest rates are bounded at zero. *The Review of Economics and Statistics* 79: 573–585.
Hicks, J. 1937. Mr. Keynes and the 'Classics': A suggested interpretation. *Econometrica* 5: 147–159.
Ireland, P. 2004. Money's role in the monetary business cycle. *Journal of Money, Credit, and Banking* 36: 969–983.
Kashyap, A., and J. Stein. 1994. Monetary policy and bank lending. In *Monetary policy*, ed. N. Mankiw. Chicago: University of Chicago Press.
Kerr, W., and R. King. 1996. Limits on interest rate rules in the IS model. *Federal Reserve Bank of Richmond Economic Quarterly* 82: 47–75.
Kimball, M. 1995. The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking* 27: 1241–1277.
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. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Lettau, M., and S. Ludvigson. 2004. Understanding trend and cycle in asset values: Reevaluating the wealth effect on consumption. *American Economic Review* 94: 276–299.
Lucas, R. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Mankiw, N., and D. Romer. 1991. *New keynesian economics. volume 1: Imperfect competition and sticky prices. volume 2: Coordination failures and real rigidities*. Cambridge, MA: MIT Press.
McCallum, B. 1979. The current state of the policy-ineffectiveness debate. *American Economic Review* 69: 240–245.
McCallum, B., and E. Nelson. 1999. An optimizing IS–LM specification for monetary policy and business cycle analysis. *Journal of Money, Credit, and Banking* 31: 296–316.
Meltzer, A. 1995. Monetary, credit and (other) transmission processes: A monetarist perspective. *Journal of Economic Perspectives* 9: 49–72.
Mishkin, F. 1995. Symposium on the monetary transmission mechanism. *Journal of Economic Perspectives* 9(4): 3–10.
Mundell, R. 1963. Capital mobility and stabilization policy under fixed and flexible exchange rates. *The Canadian Journal of Economics and Political Science* 29: 475–485.
Obstfeld, M., and K. Rogoff. 1995. Exchange rate dynamics redux. *Journal of Political Economy* 103: 624–660.
Phelps, E., and J. Taylor. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85: 163–190.
Poole, W. 1970. Optimal choice of monetary policy instruments in a simple stochastic macro model. *Quarterly Journal of Economics* 84: 197–216.
Roosa, R. 1951. Interest rates and the central bank. In *Money, trade, and economic growth: Essays in honor of John Henry Williams*. New York: Macmillan.
Rotemberg, J. 1982. Sticky prices in the United States. *Journal of Political Economy* 90: 1187–1211.
Sargent, T., and N. Wallace. 1975. 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83: 241–254.
Summers, L. 1991. How should long-term monetary policy be determined? *Journal of Money, Credit, and Banking* 23: 625–631.
Svensson, L. 2003. Escaping from a liquidity trap and deflation: The foolproof way and others. *Journal of Economic Perspectives* 17(4): 145–166.
Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1: 15–29.
Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

Money

James Tobin

Keywords
Silver standard; Numeraire; Free banking

JEL Classification
E4

Money as a Social Institution and Public Good

Among the conventions of almost every human society of historical record has been the use of *money*, that is, particular commodities or tokens as measures of value and media of exchange in economic transactions. Somehow the members of a society agree on what will be acceptable tender in making payments and settling debts among themselves. General agreement to the convention, not the particular media agreed upon, is the source of money's immense value to the society. In this respect money is similar to language, standard time, or the convention designating the side of the road for passing.

The reason for the universality of money as a social institution is that it facilitates trade. Trade among individuals enables them to achieve much higher standards of living than if each person or family were restricted to autarchic subsistence. Because of economies of scale, division of labour among specialists yields enormous gains. Of course, trades have always taken place by barter, and even in modern economies many exchanges occur without money. Barter is usually bilateral, thus in Jevons's famous phrase it requires 'a double coincidence [of wants], which will rarely happen' (1875: 3). Multilateral trade is much more efficient, permitting each trader bilateral imbalances provided her trade in aggregate is balanced. Imagine, for example, that for lack of double coincidences no bilateral trades are possible among A, B and C because A wants C's goods, B wants A's and C wants B's. Obviously three-way exchange would benefit everyone.

Multilateral barter is conceivable. It could be arranged by putting participants in simultaneous communication with each other – in person as at a village market or a commodity or stock exchange, or by modern telecommunications. But any multi-participant multi-commodity market would need a clearing mechanism. A trader would not have to be balanced with every other trader. But in the absence of a money each trader would have to be balanced in every commodity. This would be awkward and inefficient. Participants would need to come to market with inventories of many goods. A natural conclusion of any one market session would be intertemporal deals, commodities acquired today in exchange for promised future deliveries of the same or other commodities. Without money, this too would be awkward: a typical trader would end up with debts to or claims on other traders in many specific commodities.

One could imagine using intrinsically valueless tokens during a market session to lubricate barter – like poker chips for scorekeeping in a stakeless poker game. The tokens would make it possible to price each commodity in a common *numéraire* rather than in each of numerous other commodities. But if the tokens became worthless at the end of the session, each participant would have to be required to return as many tokens as he or she started with. Otherwise no one would sell useful goods for tokens, for fear of leaving the market with them rather than with commodities of value. If instead the tokens will be acceptable tenders in this and other markets in future – well, then they are money (on these issues see Hawtrey 1927, ch. 1; Starr 1972; Shubik 1984; Kareken and Wallace 1980).

The social convention makes a society's money generally acceptable within it, and the practice of general acceptability reinforces the convention. Y accepts money from X in exchange for goods and services and other things of value because Y is confident that Z, A, B,. . ., and indeed X will in turn accept that same money. Moreover, money is accepted from the bearer immediately and impersonally – without delay, without identification. Since an economic agent's purchases and sales, outlays and receipts, are not perfectly synchronous, each agent's inventory of money fluctuates in size as money circulates throughout the economy. These fluctuations in individual money holdings enable essential intertemporal exchanges to take place. Workers are paid for their labour today, and next week they buy the food and clothing that are the truly desired proceeds of their work. The farmer and the tailor accumulate money from those sales; on payday they pay it out to their hired hands.

The moneys chosen by societies have varied tremendously over human history. So have their languages. In each case, what is universal and

important is that something is chosen, not what is chosen. The variety of choice defies generalizations about the intrinsic properties of moneys. Livestock, salt, glass beads and seashells have served as money. Major grain crops were natural media for payments of wages and rents, and therefore in other transactions and accounts. Cigarettes were money in prisoner-of-war camps. On the island of Yap debts were settled by changing the ownership of large immovable stone wheels. The practice continued after the sea flooded their site and the stones were invisible at the bottom of a lagoon. (Similarly when gold was international money in the twentieth-century title to it often changed while the gold itself, safe in underground vaults, never moved.)

Some moneys have been commodities valued independently of their monetary role, intrinsically useful in production or consumption. Others have been tokens of no intrinsic utility and negligible cost of production, coins or pieces of paper. Commodity moneys derive their value partly, and token moneys wholly, from the social convention that designates them as money.

In modern nation-states the sovereign government can generally determine the society's money. For example, the United States constitution assigns to the federal government (thus, not to the states) the power 'to coin money, regulate the value thereof, and of foreign coin'. The central government defines the monetary unit, decides in what media taxes and other debts to the government itself may be paid, and defines what media are legal tender in the settlement of other debts and contracts (Starr 1974).

Precious Metals as Money

Gold and silver have histories going back many centuries as the moneys of choice of many societies and as international media of exchange. Copper coinage antedates them, but copper became too abundant and was relegated to subsidiary coins. The precious metals are durable. They are divisible into convenient denominations. They can be made into ingots, bars and coins of standard weights. When used as moneys, they have been sufficiently scarce – relative to the non-monetary demands for them – as to pack considerable value into convenient portable forms. They glitter. They have long been prized for ornament and display. Gold and silver, one or the other or both, were the basic moneys of Europe and of European dominions and settlements throughout the world from the seventeenth century, or before, until recently. In modern times gold, in particular, acquired awesome mystique (Keynes 1930).

Sovereigns minted these precious metals on demand into coins of their own realms, with their own names. In addition to minting *full-bodied* coins for public circulation, sovereigns commonly provided *token* coins made of metals, convenient for retail transactions, negligible in intrinsic value but convertible into the basic money of the realm. Many full-bodied coins circulated across national boundaries with values equivalent to their weight. For example, the original monetary unit of the United States was the silver dollar of Spanish America.

Until the late nineteenth century silver was more prevalent than gold as a monetary commodity. From medieval times silver was the English money of account; the pound sterling was initially a weight of silver. England and many other countries coined both silver and gold, but there were frequent periods when bimetallism degenerated *de facto* into one standard or the other. This happened when their prices at the mint diverged enough from their relative values in other countries or in commerce to offset the costs of arbitrage. Then 'Gresham's law' would take over, and the metal undervalued at the mint, the 'good money', would disappear from monetary circulation, 'driven out' by the 'bad money' overpriced at the mint (Hawtrey 1927: 202–4, 283).

In England in 1717 Isaac Newton, Master of the Mint, unintentionally overvalued gold, pushing silver out of circulation and in effect putting England on a gold standard. The switch was formalized in 1816. During the nineteenth century other European countries and the United States likewise gravitated from bimetallism to gold. Alexander Hamilton, America's first Secretary of the Treasury, complemented the silver dollar with

gold coins. But it was not until the late nineteenth century that gold overtook silver as the basic money of the United States. The values of sterling and dollars in gold set by Newton and Hamilton, implying an exchange rate of $4.86 per pound, lasted until 1931, with several wartime interruptions.

The heyday of the international gold standard was 1880–1914, when all major national currencies were convertible into gold at fixed rates. Silver, like copper before it, was eventually demoted to token coin status (Hawtrey 1927, chs 16–20).

Functions of Money

A triad long familiar to students of introductory economics lists the functions of money: (1) unit of account, or *numéraire*, (2) means of payment, or medium of exchange, and (3) store of value.

The US dollar, for example, is the unit of account in the United States. Prices of everything are quoted in dollars, and accounts are kept in dollars. The various media that change hands in transactions – coins, paper currency, deposits – are denominated in dollars. That does not prevent anyone who cares to do so from quoting prices in a foreign currency or in bushels of wheat, or from finding sellers who will accept them in payment for other things. It just would not be very efficient as a general practice.

To be sure, some societies have used, and kept accounts in, more than one money – in both gold and silver or, for example, in Japan two centuries ago, both in coins and in standard weights of rice. Today some national currencies may be acceptable means of payment in other jurisdictions – dollars in Russia, Israel and Canada, yen in Hawaii, Deutschemarks in Eastern Europe. The reason may be the frequency of cross-border tourism and trade. Or it may be that as a consequence of hyperinflation people turn to a 'hard' foreign currency as unit of account. For still a different reason, a new European currency, the ecu, may become a *numéraire* parallel to national currencies like pounds, francs and Deutschemarks during the period of transition to a common currency.

A society's money is necessarily a store of value. Otherwise it could not be an acceptable means of payment. (New York subway tokens cannot be generally acceptable money; they can become valueless any day, even for use as subway fare. US food stamps, intended to be in-kind welfare benefits, are exchanged with cash at par, while grocery brands' discount coupons are disqualified by their expiration dates.)

Money is the principal means of payment of a society, but it is only one of many stores of value – and quantitatively a minor one at that. Through most of human history land has been the major form of wealth, increasingly augmented by livestock and reproducible capital – buildings, tools, machines and durable goods of all kinds. Claims to much of this wealth today take the form of bonds and shares and other securities. In the United States, basic money is only 6% of total privately owned wealth.

Even though a particular commodity or token is established as the generally acceptable medium for discharging debts denominated in the unit of account, it need not be and generally is not the sole means of payment in use. *Derivative* media, often termed *representative* money, arise and circulate as media of exchange. They are promises to pay the *basic*, sometimes called *definitive*, money on demand. In the commercial city states of northern Italy, merchants left gold with goldsmiths for safekeeping. They then found it convenient to circulate the 'warehouse' receipts in place of the gold. Those payable to bearers were precursors of paper currency and banknotes. Those payable to named persons, and on their order to third parties, were precursors of cheques. Indeed, once the goldsmiths realized that they need not keep 100% gold reserves against the outstanding claims upon them, and that they could lend their certificates to merchants promising to deliver gold later, they became banks.

Besides providing token coins, states issued paper currency redeemable in gold or silver, or delegated the privilege to a private bank chartered to serve the state, like the Bank of England, founded in 1694. In addition, ordinary private banks issued their own notes, backed only by their own promises to pay basic money, gold or

silver. In the nineteenth and twentieth centuries, governments and their central banks came to monopolize the issue of paper currency. This was not a catastrophe for banks. In modern economies, demand deposits in banks, transferable to third parties by cheque or wire or other order, have become the most important derivative media of exchange.

Whether derivative moneys were officially or privately issued, the ability of the issuers to carry out their promises to redeem them in basic money, gold or silver, was a recurrent problem. In wars and other emergencies governments often suspended these promises and issued irredeemable paper money. The trend in the twentieth century was to dispense with commodity money and to replace it with fiat money of no intrinsic value. Within each nation, the official derivative money, government currency, became the basic money. In 1933 United States paper dollars became inconvertible into gold except by foreign governments or central banks.

Internationally, gold was dethroned in 1971 as the medium for settlement of imbalances of payments between countries. Governments are no longer prepared to buy or sell gold at prices fixed in their own currencies. Gold is traded freely in private markets all over the world. Its price fluctuates as people speculate about its future. In the United States there is still an official weight of gold that theoretically corresponds to the dollar – 0.0231 oz, that is a gold price of $43.22, about one eighth of the free market price. But the US government is not prepared to sell any gold for dollars at the official price – or at the free market price, for that matter.

The US monetary base (M0) is the amount of fiat currency the government, mainly its central bank, the Federal Reserve System, has issued. It is a 'debt' to the public on which the government pays no interest and against which the government holds virtually no assets (other than its remaining gold stock, $11 billion at the official price, and its drawing rights at the International Monetary Fund, $19 billion). Derivative promises to pay dollars are now, directly or indirectly, commitments to pay this fiat money. Those promises include bank deposits and all other debts, private and public, denominated in dollars and payable at specified future times, tomorrow or 30 years hence.

In the United States in the fourth quarter of 1991 the stock of *transactions money* (M1) held by economic agents other than the federal government and banks averaged $890 billion, $265 of currency (paper and coin) and $617 of chequable deposits available on demand. The banks held reserves of $53 billion in currency in their vaults or on deposit in the 12 Federal Reserve Banks, collectively the American central bank. The sum of the currency in public circulation and the currency or equivalent held as bank reserves is the *monetary base* (M0), $318 billion. It is often called *high-powered* money: every dollar of M0 was supporting $2.80 of Ml, and GNP transactions of $18.20 a year.

Sovereigns have long profited from their money monopolies. Their mints charged 'seigniorage' fees – and sometimes they cheated. Likewise, issue of currency bearing zero interest is a way for a government to pay its bills, easier than taxation and cheaper than interest-bearing debt. By regularly issuing base money to keep up with economic growth and inflation, the sovereign collects seigniorage year after year. In the United States today seigniorage is a minor source of revenue. Since base money is only 6% of GNP, growth of dollar GNP at 7% a year means new issue of base money of only 0.42% of GNP, 1.68% of the federal budget. But for many less developed countries printing money is a major way of financing public expenditures; seigniorage is a major source of revenue, because implicit taxation by inflation is politically easier than explicit taxation.

Commodity Money vs Fiat Money

The age of fiat money, first in one nation after another and finally internationally as well, has been more inflationary than the century of silver and gold standards between the Napoleonic wars and the First World War. During and following the 1914–1918 war the gold standard broke down, and attempts to re-establish it during the Great Depression did not succeed. The Bretton Woods

regime established in 1945 linked the world's currencies to gold via their fixed parities with the US dollar, because foreign governments could convert dollars into gold at a fixed price. But this system differed radically from the pre-1914 gold standard in that currency exchange rates could be and were frequently changed. The discipline imposed on a government and economy by an exchange parity fixed for a long time was diluted. In 1971, when this discipline became too much for the US itself, the gold–dollar parity gave way, and the international monetary system was wholly a regime of fiat money.

Discontent with inflation since the Second World War, and with the volatility of currency exchange rates since 1971, has led to agitation for return to the gold standard or some other commodity money. A commodity standard, if adhered to, provides a real anchor for nominal prices; its discipline prevents hyperinflation.

However, although the long-run trend of prices during the gold standard period was flat, there were violent inflationary and deflationary fluctuations around it. More important, real economic activity was highly volatile, to a degree that would be politically unacceptable nowadays (Cooper 1982, 1991).

Irving Fisher, writing during the gold standard era, was greatly concerned by the instability of prices. He was complaining, in effect, about the volatility of the relative price of gold. Ideally, he would define the dollar in terms of a representative package of goods and services, the bundle priced in a comprehensive index number. Thus he revived the idea of a 'tabular standard', proposed by several early-nineteenth-century writers, and described with approval by Jevons (1875, ch. 25). But exchange between paper currency and such bundles is impractical. Fisher proposed instead to make periodic adjustments of the gold content of the dollar, raising or lowering it in proportion to the rise or fall in the price index since the previous adjustment. In effect, the Treasury would be selling gold for dollars to fight inflation and buying gold for dollars to fight deflation (Fisher 1920).

A recent proposal by Robert Hall (1982) would tie the dollar to a composite commodity 'ANCAP' of ammonium nitrate, copper, aluminium and plywood.

Because ANCAP's prices have historically mirrored general indices, it is meant to be a feasible proxy for the economy's aggregate market basket (other proposals for commodity standards are described in Cooper 1991).

The Fisher strategy could be followed, even imposed as a nondiscretionary rule on the central bank, in a regime of fiat money. The market operations to implement it would be carried out in securities rather than in gold. The fundamental issue is not the monetary standard but whether stabilizing a price index should be the exclusive objective of monetary policy, to the exclusion of stabilization of real output growth and employment.

Free Market Money?

Would it be possible to privatize money? Certainly it is possible to privatize derivative issues of money, promises to pay fixed amounts of base money on demand. But United States experience suggests that the supply of money, even derivative 'low-powered' money, cannot safely be left to free market competition.

Before the establishment of the national banking system in 1864, private banknotes were the only paper currency of the United States. The several states freely chartered banks, and those banks freely issued their own banknotes. These were promises to pay silver dollars, but so-called 'wildcat' banks contrived to make it tough for noteholders to find them. There was no central bank to control the aggregate issue of banknotes. The notes circulated at varying discounts from par and often became worthless, stranding innocent holders.

As a result, Congress established a system of nationally chartered banks in 1864, and taxed state banknotes out of existence. Only nationally chartered banks could issue notes, and these had to be fully backed by US Treasury debt securities. In effect, they were Treasury currency, supplementing various direct issues of Treasury currency (including the inconvertible

'greenbacks' the union government issued during the 1861–1865 Civil War, which were made convertible into specie in 1879). Central banking did not begin in the United States until the Federal Reserve Act of 1914, which confined the issue of banknotes to Federal Reserve Banks.

Although private banks, state and national, were out of the business of issuing demand notes, they were still in the business of accepting demand deposits, the increasingly prevalent form of derivative money. Banks' balance sheets were regulated, but depositors were at risk. Their banks might not be able to pay in gold or equivalent on demand. After the epidemic bank failures of the 1920s and 1930s, Congress initiated a system of federal deposit insurance. Deposits in banks and other financial institutions became governmentally guaranteed, like banknotes after 1864. In the 1980s, these deposit guarantees became an expensive burden on federal taxpayers.

Could government get out of the money business altogether? It seems barely possible with commodity money and not possible with fiat money. If the government defined the *dollar* as a certain weight of gold or ANCAP or some other commodity or bundle, then private entrepreneurs could issue 'dollars', either chequable deposits or paper notes. They would be promises to pay the bearer the equivalent in the chosen commodities. The commodities themselves would not necessarily circulate on their own; indeed ANCAP and other composites could not.

The money entrepreneurs would have to keep inventories of the commodity as reserves. If one hundred per cent reserves were required, the currency would be like goldsmiths' warehouse receipts, and the private issuers would earn just a small fee for 'minting' the commodity into paper. Left to themselves, they would become banks, acquiring risky and illiquid assets while incurring demand liabilities. *Caveat emptor* would reign. The rates various banks would have to pay to attract funds would reflect depositors' appraisals of the risks. Notes and cheques of risky banks would not be honoured at par. In short, the very problems that resulted in consensus that issue of money cannot safely be left to unregulated free markets would recur.

Could the government's role be confined to defining the unit of account, the commodity equivalent of a dollar, in the same way that the government – through the Bureau of Standards in the United States – defines weights and measures? Could the system operate without any government-owned or government-issued base money? In its absence, clearings among private banks would require awkward transfers of ownership of the commodities kept as reserves against their liabilities. Very probably some one bank or consortium would arise as an unofficial central bank, and its liabilities would play the role of base money, the medium in which clearing imbalances among other banks are settled. The central bank, official or unofficial, would have to hold inventories of the standard commodity, gold or ANCAP or whatever, and be prepared to convert currency into the commodity and vice versa. That institution, history also suggests, would eventually be nationalized.

A fortiori, if there is neither an official definition of the 'dollar' nor any issue of dollars by the government or a quasi-governmental institution, there would be no standard commodity for private banks to compete in supplying to the public. Barter trading would be the rule, and the public-good advantages of social agreement on money would be lost. Since the institution of money is a public good, it is not surprising that its advantages cannot be realized by private market competition unassisted and uncontrolled.

How Can Money Have Positive Value in Exchange?

Economists have long regarded the theory of value as the central question of their discipline. What determines the prices at which goods and services are traded for each other? The prices in question include the wages of labour in terms of consumer goods, the rent of land in terms of its produce, and many other relative prices. They encompass interest rates and asset prices, thus the terms of trade of commodities to be delivered in future for commodities available today. They cover interregional and international trade, where

the prices of concern are the terms on which imports can be obtained by exports.

Money, however, is an embarrassment to value theory. According to standard theory, something can have positive value only if it generates positive marginal utility in individuals' consumption or positive marginal productivity in the making of goods and services that do generate marginal utility. The embarrassing puzzle is sharpest for fiat money. All of its value comes from the fiat that makes it money. Fiat money has no intrinsic non-monetary source of value. It cannot be eaten or worn or be used in any other way that generates utility for consumers, except a few numismatists. Nor can it contribute to the production of things that consumers do value. It can be produced at zero social cost. Yet it is a scarce commodity for any individual agent. Why is it worth anything at all? That the institution of money is of value to the society as a whole as a public good does not automatically give it value to individuals in market exchanges.

The uphill struggle of modern economic theorists to cope with these challenges is exhibited in the proceedings of a recent conference (Kareken and Wallace 1980). Their solutions relied principally on the overlapping generations model, which unrealistically assigns to money the function of being the sole or the principal store of value that links one generation to the next. The most careful, thoughtful and perceptive formal models of the roles of credit and money in transactions and strategies, in partial equilibrium and general equilibrium systems, are those of the game theorist Martin Shubik (1984).

It was argued at the beginning that a condition for fiat money to be held and valued today is that it will be acceptable in exchange for intrinsically useful commodities tomorrow. But this bootstrap story may not work. Suppose the world itself is known to be finite; its end will come at a definite future time. In the last period, 1 min before midnight so to speak, you may need money to buy whatever consumer goods might generate utility, at least solace. Otherwise you will be confined to your own resources. But who will sell you anything, knowing that the money will be worthless while the goods might be a source of some utility? Thus money is worthless 1 min before midnight, and by iterations of the same argument, it is worthless today. Even if the institution of money had public-good value between now and the end of the world, the money itself would have no market value to individuals.

The escape from this logical impasse is that we do not all and will not all expect with certainty the end of the world at any definite time. We always do, always will, assign some probability to its continuation. Since there are many other paradoxes involved in thinking about human behaviour in a world with no chance of a future beyond a definite time, it is best not to take that prospect seriously in economic modelling.

Formal general equilibrium theory, which describes the imaginary world of frictionless barter, does of course express the prices of goods and services in a *numéraire*. It is tempting to identify *numéraire* prices as money prices. But the *numéraire* is just a mathematical normalization convenient for handling the fact that the supply equals-demand equations for N commodities determine only the $N-1$ relative prices. Those relative prices are, by construction, independent of the scalar arbitrarily attached to the *numéraire*.

Standard value theory does, of course, have something to say about the value of commodity money in terms of other goods and services. In a gold standard regime, the relative prices of gold in other commodities have to be the same at the mint and in the market; they cannot depend on whether the gold is circulating in coins or being used in jewellery, dentistry or rocketry. That is simply a condition of the absence of arbitrage profits. It definitely does not say that under the gold standard the relative price of gold is the same as it would be if gold were not money. As argued above, gold's role as money must increase the demand for it, and that must affect its price unless it is supplied perfectly elastically. The same will be true of any other commodity or bundle of commodities chosen as the monetary standard. A substantial part of the value of any commodity used as money arises from the convention or the fiat that makes it money. The distinction between commodity money and fiat money is not absolute.

The Neutrality of Money

Although business managers, financiers, politicians and workers worry a great deal about monetary institutions and policies and their consequences for economic activity and well-being, pure economic theory minimizes these consequences. Theory puts the burden of proof on anyone who contends that money and monetary inflations or deflations do much good or much ill.

Classical economists liked to insist that money is a veil, obscuring but not altering the real economic scenario (Robertson [1922] 1959:7). Their modern descendants expound 'real business cycle theory', premised on the view that economic developments that matter to societies and individuals are independent of monetary events and policies (Prescott 1986). It is true that economic fluctuations and trends are frequently misinterpreted by stressing superficial monetary phenomena to the neglect of resources, technologies and tastes. But money does matter, really.

Does an economy arrive at the same *real* outcomes (in variables like volumes of production, consumption and employment, and in relative prices such as the purchasing power of wages and the price of oil relative to that of bread) as it would without the institution of money? Clearly not. Without money, confined to barter, the economy would produce a different menu of products, less of most things. People would spend more time searching for trades and less in actual production, consumption and leisure.

That is not the comparison the classical economists, old and new, intend by the 'veil' metaphor. Their fantasy is a frictionless, costless system of multilateral barter, in which relative prices and the allocations of labour and capital among various productive activities are determined in competitive markets. Their proposition is that the outcomes of an economy with money are the same as those that would arise from their ideal barter model. The corollary is that real economic outcomes are independent of the particular nature of the monetary institutions (Dillard 1988).

These propositions cannot be true of commodity money. Real economic outcomes with commodity money will differ from those with fiat money, and will also depend on what commodity is selected as money. Inventories of the chosen commodity have to be held for exchange purposes and for governmental and bank reserves, beyond the stocks held in connection with the commodity's non-monetary uses in production and consumption. In growing economies demands for monetary inventories will be steadily increasing. The relative demands for monetary and non-monetary inventories are bound to change with economic and technological developments that alter the incentives to produce the commodity and change its prices in terms of other goods and services. Examples are discoveries or exhaustions of gold and silver deposits and innovations in mining and processing technologies. Since the monetary commodity's price is fixed in money, its output will decline when there is general inflation and rise when there is deflation. Intertemporal choices involving the monetary commodity, as well as contemporaneous choices, will be significantly affected by its monetary use.

The availability of moneys, whether commodity or fiat, whether basic or derivative, as stores of value necessarily brings about significant deviations in real outcomes from the hypothetical regime of frictionless barter. This is true even though that regime is postulated to include markets in state-contingent commodity futures, 'Arrow–Debreu' contracts (Arrow and Debreu 1954). Holding monetary assets gives agents more flexibility: they can convert them into consumption of any kind at any time in any 'state of nature', though not at predictable prices. The flexibility is a convenience to individual agents. But, as Keynes saw, it opens the door to 'coordination failures' which are the essence of macroeconomics – demand for goods and services may at times diverge seriously from supplies (Keynes 1936, chs 16, 17).

The Classical Dichotomy

It is possible to recognize that an economy with monetary institutions is different in real outcomes from a barter economy, even from an ideal frictionless barter economy, and still to argue that its

real outcomes are independent of the purely nominal parameters of those institutions. It would be terribly convenient if the determination of the absolute price level, the reciprocal of the value of the monetary unit in a representative bundle of consumer goods, could be split off from the determination of relative prices and the associated real quantities.

Don Patinkin (1956) called this separation the *classical dichotomy*. Only monetary shocks would affect the general price level, and those shocks would raise or lower the nominal prices of all commodities in the same proportions. Only real shocks – to tastes, technologies and resource supplies – would affect relative prices and real quantities. This proposition would not exclude the fact that the monetary institutions themselves matter. The choice between commodity money and fiat money, the choice among possible commodity standards, and the arrangements for derivative moneys might well affect the social efficiency of markets and trade.

What are the nominal parameters whose settings, according to the classical dichotomy, would make no real difference? For a commodity money, such a parameter is the definition of the monetary unit in terms of the standard commodity, for example the weight in gold of a dollar. For fiat money, the key nominal parameter is the quantity of money – base money, all transactions money, or some even more inclusive aggregate.

Why should cutting the gold content of the dollar from 0.0484 ounces to 0.0286 ounces, raising the dollar price of gold from $20.67 to $35.00 (as Franklin Roosevelt did in 1933), make any real difference? The dollar values of existing public and private stocks of gold, and of monetary claims to gold would rise in the same proportion. Will not all other commodity prices do likewise? Then all relative prices and real quantities, including those of gold, will be the same as before.

For fiat money systems, and for commodity standards where issues of derivative moneys have become essentially independent of the commodity, the *quantity theory of money* achieved similar dichotomization. According to the theory, which might more accurately be called the quantity-of-money theory of prices, an increase in the nominal quantity of money would raise all nominal commodity prices in the same proportion, leaving relative prices and real quantities unchanged. Quantity theorists argue that an increase in the quantity of money is equivalent to a change in the monetary unit. A 100-fold increase in the stock of French francs would be – would it not? – the same as De Gaulle's decree changing the unit of account to a new franc equivalent to 100 old francs. Since the units change could make no real difference, the other way of multiplying the money stock could not either.

These analogies fail, for several related reasons. In most economies money is by no means the only asset denominated in the monetary unit. There are many promises to pay base money on demand or at specified dates. If there is a thorough units change, like De Gaulle's, all these assets are automatically converted to the new unit of account. Roosevelt's devaluation of the dollar relative to gold was not a pure units change. He did not scale up the dollar values of outstanding currency or even of Treasury bonds with provisions for such revaluation. Naturally private assets and debts expressed in dollars were not scaled up either. Likewise, when the quantity of money is changed by normal operations of governments or central banks or by other events, the outstanding amounts of other nominally denominated assets are not scaled up or down in the same proportion. They may remain constant, as when money is printed to finance government expenditures. They may move in the opposite direction, as when central banks engage in open-market operations, which typically increase the amount of base money outstanding by buying bills or bonds, thus reducing the quantities of them in the hands of the public.

The Quantity Theory

The quantity theory goes back to David Hume, probably farther, but its major and most effective protagonists have been Irving Fisher (1911) and Milton Friedman (1956).

In its crudest form, the quantity theory is a mechanistic proposition strangely alien to the

assumptions of rational maximizing behaviour on which classical and neoclassical economic theories generally rely, as J.R. Hicks eloquently pointed out in a famous article (1935). Specifically, it ignores the effects of the returns to holding money on the amounts economic agents choose to hold. The technology of monetary circulation fixes the annual turnover of a unit of money. Suppose that every dollar 'sitting' supports just V dollars per year 'on the wing', to use D.H. Robertson's famous terms ([1922] 1959: 30). Suppose, further, that the economy is assumed to be in real equilibrium and the supply of money is doubled. The public will not wish to hold the additional money until the dollar value of transactions is doubled, and this requires prices to double.

Surely the demand for money to hold is not so mechanical. The velocity of money can be speeded up if people put up with more inconvenience and risk more illiquidity in managing their transactions. Money holdings depend, therefore, on the opportunity costs, the expected changes in the value of money and the real yields of other assets into which the same funds could be placed. Fisher and Friedman would agree.

The quantity theory can still be rationalized, as a proposition in comparative statics. Compare, for example, two stationary situations of a given economy, in each of which the money supply and price level are constant over time. Let the money supply in the second situation be twice that in the first. Then an equilibrium in the second situation will be the equilibrium of the first with a nominal price level twice as high. This will be true even if the demand for money is modelled as behavioural, not mechanical, and is allowed to depend on interest rates, expected inflation and other variables.

However, it is not sufficient to double solely the quantity of money, narrowly defined. All exogenous nominal quantities, including outstanding stocks of debts and assets, must also be doubled. Or the second equilibrium must be interpreted as a stationary state that will be reached only when all these other nominal stocks have had time to adjust endogenously to the new quantity of money. This quantity theory does not apply to short-run changes in monetary quantities engineered by central banks, for the same reasons that render the 'units change' metaphor inapplicable.

In its interpretation as a proposition in long-run comparative statics, the quantity theory supports 'neutrality' as asserted in the classical dichotomy. Neutrality has come to have two meanings in monetary economics. Simple *neutrality* means that real economic outcomes are independent of the levels of nominal prices. *Superneutrality* means that those outcomes are also independent of the rates of change of nominal prices.

The case for superneutrality appeals to, and depends upon, the 'Fisher equation'. Early on, Fisher (1896) saw the importance of distinguishing between nominal and real rates of interest on assets and debts denominated in monetary units. *Ex post*, the algebraic difference between them is by definition the rate of inflation or deflation. This is a tautology. But Fisher (1911) is also credited with a meaningful proposition: anticipation of inflation (deflation) raises (lowers) nominal rates of interest but does not alter real rates of interest. The corollary is that whatever is the time path of money stocks that determines the path of prices, the paths of real economic variables are the same. Fisher himself was enough of a classical economist to believe this as a long-run theoretical truth, but enough of a pragmatic empiricist to find that nominal rates were very slow to incorporate adjustments for ongoing inflations and deflations.

The Price of Money

A 1975 conference on monetarism at Brown University is remembered for a pithy observation by Milton Friedman, offered only half in jest:

> For the monetarist/non-monetarist dichotomy, I suspect that the simplest litmus test would be the conditioned reflex to the question, 'What is the price of money?' The monetarist will answer, 'The inverse of the price level'; the non-monetarist (Keynesian or central banker) will answer, 'the interest rate'. The key difference is whether the stress is on money viewed as an asset with special characteristics, or on credit and credit markets, which leads to the analysis of monetary policy and

> monetary change operating through organized 'money', i.e. 'credit', markets, rather than through actual and desired cash balances. Though not so obvious, the answer given also affects attitudes toward prices: whether their adjustment is regarded as an integral part of the economic process analyzed, or as an institutional datum to which the rest of the system will adjust (Stein 1976: 316).

'What am I', asked the chairman of the session, George Borts, 'if I answer "one"?'

Any durable good has at least two 'prices', the price at which it can be bought or sold, and the price of the services it renders per unit time. The price of the good itself is the present value of the expected, though uncertain, values of the services it will render in future. For money, the first price is its purchasing power. Its services come in two forms: as a store of value, the capital gain or loss from changes in its purchasing power, and, as a medium of exchange, the benefits it yields in convenience, effort-saving and risk reduction. Without cash on hand, an economic agent may find it costly to make desirable transactions, or to forgo them. The marginal productivity of holding money is the value of an additional dollar in reducing those costs.

What is the marginal opportunity cost to which agents will equate the marginal productivity of holding money? It depends on what alternatives are available. If money proper were the only store of value in the economy, the opportunity cost of holding money would be the marginal utility of immediate consumption relative to future consumption. Although this set-up is all too common in the literature, it confuses theories of money and of saving. Acknowledging the availability of other stores of value makes the cost of holding money the difference between the real capital gain or loss on money and the real rate of return on the non-money assets in which a marginal dollar could be invested.

If money proper were the only store of value in the monetary unit of account, though not the only one in the economy at large, the relevant opportunity cost would be the return on real capital, that is storable or durable commodities. In modern economies, however, the immediate substitutes for money are promises to pay money in future. Since money and these substitutes are affected equally by price level changes, the opportunity cost is simply the nominal interest rate on those non-money substitutes. (This assumes zero nominal interest on money itself.)

Friedman's Keynesian is careless if he calls any of these opportunity cost concepts the price of money. These are prices of the services of money. Friedman's monetarist is right, therefore, to say that the price of money is the reciprocal of the commodity price level – the real price, that is, for Borts was right about money's nominal price. Of course, there are as many relative prices as there are non-monetary commodities, and any average value of money requires using an arbitrary commodity price index.

To implement Friedman's asset valuation approach to the price of money, suppose that the nominal supply of money per capita, real per capita output and the real interest rate all follow arbitrary variable paths, anticipated in advance. Assume, at least for illustrative purposes, the Allais–Baumol–Tobin model of the demand for money (Baumol and Tobin 1989). The marginal productivity of nominal cash holdings for a representative agent is the reduction in the frequency and cost of exchanges back and forth between money and dollar-denominated interest-bearing substitutes. It is, by the usual approximation equal to $a(t)y(t)/(2(t)_2v(t))$, where a is the real cost of one of those exchanges, y is the agent's real income per period, m is the agent's average nominal cash holding, and v is the value of money, the reciprocal of the price level. Of these, a, y and m are arbitrary exogenous functions of time, while the valuation v is a function of time to be determined. Let $r(t)$ be the exogenous path of the real interest rate. The value of money at any time T is the discounted value of its future marginal productivities:

$$v(T) = a(T)\int_T^\infty \exp\left(-\int_T^t r(s)\mathrm{d}s\right)y(t) \Big/ \left(2m(t)^2 v(t)\mathrm{d}t\right), \tag{1}$$

$$v'(T) = r(T)v(T) - a(T)y(T) \Big/ \left(2m(T)^2 v(T)\right), \tag{2}$$

$$r(T) - v'(T)/v(T) = a(t)y(T) \Big/ \left(2m(T)^2 v(T)^2\right). \tag{3}$$

Equation (3), with the nominal interest rate on the left, is the familiar equation for optimal *real* cash holdings. It involves the stronger Fisher equation, because the real rate has been taken as exogenous.

Interpreted as the price dynamics of the economy, these equations describe the time path of the 'price of money'. The level of prices at each time converts the autonomous nominal money supply into the real quantity on which its marginal productivity depends. The price path itself generates the rates of price change which, added to the autonomous real interest rates, give the nominal rates. The marginal productivity of money at each point in time is equated to the nominal interest rate. Future as well as current values of money supplies, as well as other variables, affect current prices. An expected increase in future money supply raises prices today, and so does an expected future increase in real rates of interest. The Fisher equation is essential to maintain the assumed dichotomy between the paths of real and nominal variables (for a calculation in this same spirit, see Sargent and Wallace 1981).

Money and Macroeconomics

In the above scenario, a key institutional fact is that the nominal interest rate on money proper is fixed, at zero. Expected inflation makes money's real interest rate negative and reduces the attraction of holding money compared to assets bearing the economy's real interest rate. For the same reason, an increase in that real interest rate is a disincentive to hold money.

However, the same institution – the fixed nominal interest rate on money – threatens the classical dichotomy. It calls into question the Fisher equation, which is central to the independence from monetary influence of the real rate of interest and related real variables. It calls it into question in principle, in long runs and short, in equilibrium and in disequilibrium. If expected inflation diminishes demand for money, it by the same token increases demands for other assets, both interest-bearing promises to pay money and real capital. These substitutions will reduce the real interest rates on those assets; their nominal interest rates will rise less than the full inflation premium. This effect – associated in the literature with the names of Mundell (1963) and Tobin (1965, 1969) – refutes superneutrality, which is essential to neutrality in any general dynamic meaning. That is to say, it is not possible to determine the real interest rate and related real variables independently of the money equation, or to determine the value of money from the demand = supply equation for money by itself.

This is true whether the economy is assumed to be classical, with full employment assured by flexibility of nominal interest rates and prices, or Keynesian, with aggregate demand short of full employment. However, the real effects of expected price inflation and deflation are a reason for doubting the efficacy of price flexibility in sustaining or restoring full employment equilibrium in the face of aggregate demand shocks (Fisher 1933; Keynes 1936, ch. 19; Tobin 1975).

Irving Fisher, Alfred Marshall and other monetary economists of the early twentieth century regarded neutrality in any sense as properties of long-run static equilibrium, not of the dynamic transitions that dominate empirical observations of monetary and real variables. According to them, people are slow in translating experience of inflation into their expectations of the future. This is how Fisher interpreted the strong positive correlations he found between inflation rates and real output (Fisher 1911). However, the Mundell–Tobin effect suggests a still stronger conclusion, since it calls into question the Fisher equation even when inflation expectations are correct and people are not victims of 'money illusion'.

In Friedman's litmus test there is much more at stake than meets the eye. The issue is how the price level, whose reciprocal is the 'price of money', is determined. The monetarist's trained instinct is to think of it as determined by the demand = supply equation for money 'as an asset with special characteristics'. With the absolute price level thus determined, the function of

markets for goods and services is to generate real, relative prices, just as in Walrasian general equilibrium theory. Those real variables, in turn, are exogenous to the path of the 'price of money'.

The Keynesian's trained instinct, on the other hand, is to think of the price level as an index of nominal prices of goods and services. As Keynes (1936, Book I) emphasized – for labour markets especially – markets in our monetary economies determine in the first instance nominal prices, not real prices. The pricel 'level' is a synthetic aggregate of multitudes of individual prices determined in diverse imperfect markets, often decided by administrative decisions or by negotiations. For price determination the most relevant equations of a macroeconomic model are price and wage equations, often members of the Phillips curve family. These specify inertia of varying degrees in nominal prices and relate their changes to measures of real excess demand or supply. As a result, price indices move smoothly and sluggishly over time, not 'jumping' like the price of a financial asset sensitive to market views of the future.

With the price level determined in goods markets, the function of the money demand = supply equation is to generate interest rates. That explains the Keynesian's instinctive response to the test question. Of course, the Keynesian recognizes that the endogenous variables of a simultaneous equations system are determined jointly, not equation by equation. That real variables are among those endogenous variables can be attributed to the fact that there is usually a non-zero discrepancy between the price path determined by the full system and the path that would be generated by the monetarist's asset price of money. The non-monetarist view does not take prices 'as an institutional datum to which the rest of the system will adjust', but it does rely on variables besides prices to equate 'actual and desired cash balances'.

The equation of money demand and supply is just one of many relations in a theoretical or econometric macroeconomic model. The small tail cannot wag the big dog. That was too much to expect. The price level is a factor common to the valuation of many assets denominated in the monetary unit, many of them close substitutes for transactions money. Their quantities now and in future must make a difference. Of course monetary policies and supplies, current and prospective, are important determinants of the price level, and so are credit markets. But the channels of these influences run through demands and supplies in markets for goods and services. Understanding the process belongs to the messy subject of macroeconomics. Finance theory, however elegant, cannot provide a shortcut.

Monetary events and policies are not a sideshow to the main performance. The real variables of a monetary economy are hopelessly entangled with monetary phenomena. They do not behave as if an economy enjoying the societal advantages of money were a frictionless multilateral barter economy seen through a veil. That barter economy would never have business cycles characterized by economy-wide excess demands and supplies of labour and other goods and services. The public-good advantages of the institution of money do not come so cheap. Among their costs are fluctuations in business activity and in the value of money itself. Pragmatic monetary economics is a central part of macroeconomics in general.

See Also

- ▶ Fiat Money
- ▶ Financial Intermediation
- ▶ Gold Standard
- ▶ International Finance
- ▶ Monetarism
- ▶ New Classical Macroeconomics
- ▶ Quantity Theory of Money
- ▶ Rational Expectations

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Baumol, W.J., and J. Tobin. 1989. The optimal cash balance proposition: Maurice Allais's priority. *Journal of Economic Literature* 27: 1160–1162.

Cooper, R.N. 1982. The gold standard: Historical facts and future prospects. *Brookings Papers on Economic Activity* 1982(1): 1–45.

Cooper, R.N. 1991. Toward an international commodity standard? In *Money, macroeconomics, and economic policy*, ed. W.C. Brainard, W.D. Nordhaus, and H.W. Watts. Cambridge, MA: MIT Press.
Dillard, D. 1988. The barter illusion in classical and neoclassical economics. *Eastern Economic Journal* 14: 299–318.
Fisher, I. 1896. *Appreciation and interest*. Publications of the American Economic Association, 3rd series 11(4); reprinted. Fairfield: A.M. Kelley, 1991.
Fisher, I. 1906. *The nature of capital and income*. New York: Macmillan.
Fisher, I. 1911. *The purchasing power of money*. New York: Macmillan.
Fisher, I. 1920. *Stabilizing the dollar*. New York: Macmillan.
Fisher, I. 1933. The debt-deflation theory of great depressions. *Econometrica* 1: 337–357.
Friedman, M. 1956. *Studies in the quantity theory of money*. Chicago: University of Chicago Press.
Hall, R.E. 1982. Explorations in the gold standard and related policies for stabilizing the dollar. In *Inflation: Causes and effects*, ed. R.E. Hall. Chicago: University of Chicago Press.
Hawtrey, R.G. 1927. *Currency and credit*, 3rd ed. London: Longmans, Green & Co.
Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica*, NS 2(1), 1–19.
Jevons, W.S. 1875. *Money and the mechanism of exchange*. London: King.
Kareken, J.H., and N. Wallace (eds.). 1980. *Models of monetary economics*. Minneapolis: Federal Reserve Bank.
Keynes, J.M. 1930. Auri sacra fames. In *Essays in persuasion*, reprinted in The collected writings of John Maynard Keynes, vol. 9. London: Macmillan. 1972; New York: Harcourt Brace.
Keynes, J.M. 1936. *The general theory of employment, interest, and money*. Reprinted in The collected writings of John Maynard Keynes, vol. 7. London: Macmillan. 1973; New York: Harcourt Brace.
Mundell, R.A. 1963. Inflation and real interest. *Journal of Political Economy* 71: 280–283.
Patinkin, D. 1956. *Money, interest, and prices*. New York: Harper and Row; 2nd edn, 1965.
Prescott, E. 1986. Theory ahead of business cycle measurement. *Federal Reserve Bank of Minneapolis Quarterly Review* 10(4): 9–22.
Robertson, D.H. 1922. *Money*, Cambridge economic handbook, 4th ed. Chicago: University of Chicago Press, 1959.
Sargent, T.J., and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3): 1–17.
Shubik, M. 1984. *A game-theoretic approach to political economy*. Cambridge, MA: MIT Press.
Starr, R.M. 1972. The structure of exchange in barter and monetary economies. *Quarterly Journal of Economics* 86: 290–302.
Starr, R.M. 1974. The price of money in a pure exchange economy with taxation. *Econometrica* 42: 45–54.
Stein, J.L. (ed.). 1976. *Monetarism*. Amsterdam: North-Holland.
Tobin, J. 1965. Money and economic growth. *Econometrica* 33: 671–684.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 1(1): 15–29.
Tobin, J. 1975. Keynesian models of recession and depression. *American Economic Review* 65: 195–202.

Money and General Equilibrium

Douglas Gale

Abstract

The study of money and general equilibrium deautility associated with the consuls with the integration of monetary theory and the classical theory of value. It includes such topics as the role of money in exchange, the determination of the price level, and the 'real' effects of money on the allocation of goods and services.

Keywords

Cash-in-advance constraint; Classical dichotomy; Complete markets; Excess demand; Financial securities; Incomplete markets; Intertemporal substitution effect; Money; Money demand; Money in general equilibrium; Money supply; Nominal prices; Optimum quantity of money; Pigou effect; Real balance effect; Say's Law; Stationary states; Temporary equilibrium; Uniform tightness property; Value theory; Walras's Law; Wealth effect

JEL Classifications

D5; E4

The general equilibrium theory of value, as developed by Walras (1874–77) and his followers, determines the relative prices of goods in terms of non-monetary factors such as technology, preferences, and endowments. Monetary factors are

used to determine the nominal price level once relative prices have been determined. Relative prices are determined by the market-clearing conditions for goods whereas the general price level is determined by the market-clearing condition for money. Given a vector of nominal prices $p = (p_1,\ldots, p_\ell)$, the market excess demand functions can be denoted by $f(p) = (f_1(p),\ldots,f_\ell(p))$, where p_h denotes the nominal price of good h and $f_h(p)$ denotes the market excess demand for good h. The functions $f(p)$ are assumed to be homogeneous of degree zero in nominal prices:

$$f(p) = f(tp),$$

for any positive scalar $t > 0$. The market-clearing conditions for goods require that the excess demand for each good vanishes at the equilibrium price vector p^*, that is $f(p^*) = 0$. These conditions can at most determine relative prices, because if p^* is an equilibrium price vector, then so is tp^*, for any positive scalar $t > 0$.

To determine the nominal price level, a demand function for money is introduced. The aggregate demand for money is assumed to be a function of prices $M(p)$. Money demand is homogeneous of degree one in prices:

$$M(tp) = tM(p),$$

for any price vector p and any scalar $t > 0$. For any vector of nominal prices p^* satisfying the goods market-clearing condition $f(p^*) = 0$, there is a unique value of $t > 0$ such that

$$M(tp^*) = \overline{M},$$

where $\overline{M} > 0$ is the exogenous money supply. Thus, once relative prices have been determined by the real factors, the level of nominal prices is determined by monetary factors. This doctrine, which became known as the classical dichotomy, characterized the classical (pre-Keynesian) thinking about monetary economics (see Fisher 1963, for example).

The integration of monetary theory and the theory of value was stimulated by the appearance of Keynes's General Theory (Keynes 1936). Pigou (1943) argued that the demand for goods could not be homogeneous of degree zero in prices, because a general fall in prices would increase the real value of money and the wealth effect would in turn increase demand for goods. The Pigou effect (the effect of a general fall in prices on the aggregate demand for goods) is a special case of the real balance effect: that is, the effect of any change in real balances on the aggregate demand for goods. In an attempt to make sense of Keynes's short period analysis, Hicks (1946) introduced the concept of temporary equilibrium, in which prices adjust to clear markets in a particular time period, taking as given expectations about prices in future periods. Building on the work of Hicks and Pigou, Patinkin (1965) argued that the real balance effect is essential for the existence and stability of equilibrium. The classical writers assumed that the market excess demand functions satisfy Say's Law, that is, the value of excess demands for goods sum to zero or

$$p \cdot f(p) = 0,$$

for any price vector p. However, Patinkin pointed out that Walras's Law should also be satisfied: that is, the value of the excess demands for goods and money should sum to zero, or

$$p \cdot f(p) + M(p) - \overline{M} = 0,$$

for any price vector p. Say's Law and Walras's Law together imply that

$$M(p) = \overline{M},$$

for any price vector p. Then homogeneity of the excess demand function $f(p)$ once again implies that, if p^* is a market-clearing price vector, so is tp^* for any $t > 0$ and the price level is once again undetermined. To avoid this indeterminacy, Patinkin argued that there must be a real balance effect: a change in the general price level implies a change in real balances, and hence a change in wealth which must change the demand for commodities. Thus, in a monetary economy the excess demand for goods $f(p,\overline{M})$ is a (homogeneous of degree zero) function of nominal prices and the money supply.

Hahn (1965) pointed out another problem in the theory of monetary equilibrium, viewed from the Walrasian perspective. The problem was the lack of a proof that money has positive value in equilibrium. Hahn observed that the uses of money that might be expected to give rise to a positive demand for money all require money to have positive value in exchange. If the value of money were zero, the economy would be identical to a barter economy. Under the usual assumptions on the excess demand functions, such a non-monetary economy would possess an equilibrium, but it would not be a monetary equilibrium, because money would have no role in exchange.

Grandmont (1983) provided an elegant solution to the problem posed by Hahn (1965). He showed that, while the real balance effect might be necessary, it was not sufficient for the existence of an equilibrium in which the value of money is positive. A strong intertemporal substitution effect is needed as well. Consider an economy in which there are two periods (the present and the future). In the first period, agents buy and sell goods for immediate consumption. They also demand money as a store of value, which they hold until the following period. The value of money is given by an indirect utility function $v(m, p')$, where $m > 0$ is the amount of money held until the future and p' is the vector of future nominal prices. An agent's expectations are represented by a probability measure μ on the space of price vectors. Expectations of future prices depend on current prices p via the expectation function $\mu = \psi(p)$. Then the expected utility associated with the cash balance m is simply the expected value of $\tilde{v}(m, p')$, conditional on the current price vector p:

$$v(m,p) = \int \tilde{v}(m,p')\mathrm{d}\psi(p).$$

Let $u(x)$ denote the utility associated with the consumption of a vector of current goods x. Then the agent seeks to maximize

$$u(x) + v(m,p)$$

subject to the budget constraint

$$p \cdot x + m \leq p \cdot e + \overline{m},$$

where e is the agent's endowment of goods and $\overline{m}$ his endowment of money. The crucial assumption (sufficient condition) for the existence of an equilibrium in which money has a positive value is that the expectation function $\psi(p)$ satisfies the uniform tightness property: for any number $\varepsilon > 0$ and for every current price vector p, there is a compact set K in the space of positive prices such that $\psi(p)$ assigns probability at least $1 - \varepsilon$ to the event that the future price vector p' belongs to K.

While the classical dichotomy cannot hold in the short run, Archibald and Lipsey (1958) argued that it would hold in the long run because the allocation of money balances is endogenous in the long run. This gave rise to the study of stationary states (see Grandmont 1983).

The Cash-in-Advance Constraint

Introduced by Clower (1967), the cash-in-advance constraint provides a simple motivation for the use of money as a medium of exchange. Lucas (1980) derives the cash-in-advance constraint as follows. Every household is assumed to consist of two agents, one of whom is responsible for selling the household's endowment of goods (for example, supplying labour) and the other is responsible for purchasing goods. At the beginning of each day, the seller sets off for the market with a bundle of goods to sell, while the buyer sets of for a different set of markets to buy the goods they need. Following Clower's dictum that 'money buys goods and goods buy money but goods do not buy goods', the buyer needs to have a stock of money at the beginning of the day. The money earned by the seller is not available until the end of the day, so the buyer's purchases are constrained by the amount of money she has at the beginning of the day. The money brought home by the seller must be held until the next day. If $\overline{m}$ is the amount of money held initially and m is the amount carried forward to the next day, the budget constraint can be written as

$$p \cdot x + m \leq p \cdot e + \overline{m}$$

and the cash-in-advance constraint can be written as

$$p \cdot (x - e)^{+} \leq \overline{m},$$

where ξ^{+} denotes the vector consisting of the non-negative part of the vector ξ.

Grandmont and Younes (1973) used a cash-in-advance constraint to study the efficiency of monetary equilibrium. They considered stationary equilibria of an infinite-horizon, pure-exchange economy in which a finite number of individuals $i = 1,\ldots, I$ maximize the discounted sum of utilities $\sum_{s=t}^{\infty} \delta^{s-t} u_i(x_i(s))$ subject to a sequence of budget constraints and a cash-in-advance constraint in the form

$$p(t) \cdot (x_i(t) - e_i)^{+} + kp \cdot (x_i(t) - e)^{-} \leq m(t-1),$$

where $0 \leq k \leq 1$. For $k = 0$ this constraint reduces to the Clower–Lucas version. Grandmont and Younes established Friedman's optimum quantity of money result: any laissez-faire, stationary equilibrium of this economy is Pareto inefficient but, if the rate of price deflation equals the subjective rate as time preference, this is sufficient to guarantee that equilibrium is efficient. Grandmont and Laroque (1975) also showed that the payment of interest on money has no effect on efficiency. More precisely, it is the gap between the inflation rate and the interest rate which has an effect, and this is attributable to the lump-sum taxes rather than the interest payments.

The cash-in-advance constraint has played an important role in macroeconomics, particularly in the study of the effect of fiscal and monetary policy (see, for example, Lucas and Stokey 1983, 1987; Sargent 1987).

Financial Securities

The classical model of general competitive equilibrium assumes that markets are complete. Hart (1975) showed that, with incomplete markets, the existence of equilibrium is no longer guaranteed and the fundamental theorems of welfare economics no longer hold. In Hart's model, incomplete markets are represented by trade in real securities, which are promises to deliver bundles of commodities at some future date and event. Cass (2006) and Werner (1985) introduced financial securities, whose payoffs are denominated in units of money, and showed that this resolved the existence problem. However, as Balasko and Cass (1989) and Geanakoplos and Mas-Colell (1989) showed, financial securities also introduced indeterminacy of equilibrium. The problem is that a change in the price level in some state changes the real purchasing power of money and hence changes the real payoffs of the financial securities. Magill and Quinzii (1992) pointed out that the indeterminacy arises from the fact that 'money' serves only as a unit of account in the Cass–Werner model. Money has no role in exchange or savings and investment, and hence there is no well defined demand for money.

To address this problem, Magill and Quinzii introduce a cash-in-advance constraint in the spirit of Clower (1967). There are two dates, $t = 0,1$, and S states of nature, $s = 1,\ldots, S$. The state is unknown at date 0; the true state is revealed at date 1. It is convenient to treat the situation at date 0 as another state, denoted $s = 0$. Then each period s is divided into three sub-periods, denoted s_1, s_2, and s_3. In subperiod s_1, agents sell their entire endowment of money to a central exchange and receive money instead. In sub-period s_2, they invest in financial securities (at date 0) and receive dividends (at date 1). In sub-period s_3, they use money to purchase goods from the central exchange. The separation of the sale and purchase of goods between sub-periods s_1 and s_3 forces agents to hold money in equilibrium. Money can also be used to store wealth between periods 0 and 1, but agents will do this only if they anticipate deflation. The supply of money is determined exogenously by the government.

Three main results were established by Magill and Quinzii. First, they showed that, *generically in endowments and money supply, an economy*

has a finite number of locally unique monetary equilibria. This means that equilibrium is locally determinate: the well-defined demand for money has eliminated the indeterminacy of the price level. Second, *if money is used as a medium of exchange only, local changes in the money supply have no real effects if the asset markets are complete* – changes in the money supply will change the price level but this will have no effect on the real allocation as long as markets are complete – *whereas, if markets are incomplete, local changes in money supply translate into an* $S - 1$ *dimensional submanifold of real allocations*. When markets are incomplete, any change in the price level implies a change in the real payoffs of the securities, and this translates into a real change in the allocation. Finally, *if money is used as a store of value, local changes in the money supply translate into an S-dimensional submanifold of real allocations in the case of both complete and incomplete markets.* This follows because the use of money as a store of value to transfer wealth between periods implies that the real allocation is directly impacted by changes in the real payoffs from holding money.

A related study by Geanakoplos and Dubey (1992) addresses a similar set of questions, but does so in the context of a model with a banking system.

Market Games

To provide microeconomic foundations for monetary equilibrium, Shubik (1972) introduced a game that integrates the use of money as a medium of exchange with a generalized Nash–Cournot model of markets. The generalization by Shapley and Shubik (1977) can be summarized as follows. There is an exchange economy with ℓ commodities, indexed by $h = 1,\ldots, \ell$, and I traders, indexed by $I = 1,\ldots,I$. Each trader is characterized by a consumption set $\mathbf{R}^{\ell}_{+}$, an endowment $e_i \in \mathbf{R}^{\ell}_{+}$, and a utility function $u_i : \mathbf{R}^{\ell}_{+} \to \mathbf{R}$. The utility functions are assumed to be C^1, nondecreasing and concave. We assume that each commodity has a positive aggregate endowment $e_h > 0$ and that each individual has a non-zero endowment $e_i > 0$.

For simplicity, we assume that traders offer their entire endowment of assets for sale and then bid for the assets they want to hold using fiat money as a means of payment. Each trader i has an endowment of fiat money $m_i > 0$. The amount of money he bids for asset h is denoted by $b_{ih} \geq 0$ and the vector consisting of his bids is denoted by $b_i \in \mathbf{R}^{\ell}_{+}$.

A trader cannot bid more money than he holds, so the bid vector chosen by trader i must satisfy the cash-in-advance constraint

$$\sum_{h=1}^{\ell} b_{ih} \leq m_i.$$

The set of bid vectors satisfying the cash-in-advance constraint for trader i is denoted by B_i, where it is understood that the initial balance m_i is exogenously given.

For any strategy profile $b = (b_1,\ldots, b_I)$, define an attainable allocation of commodities as follows. Let the price of commodity h be denoted by $p_h(b)$ and defined by

$$p_h(b) = \frac{b_h}{e_h},$$

where $b_h \equiv \sum_{i=1}^{I} b_{ih}$ and $e_h = \sum_{i=1}^{I} e_{ih}$. Then let the quantity of commodity h received by trader i be denoted by $\xi_{ih}(b)$ and defined by

$$\xi_{ih}(b) = \begin{cases} b_{ih}/p_h & \text{if } p_h > 0 \\ 0 & \text{if } p_h = 0. \end{cases}$$

Then the commodity bundle achieved by i for any strategy profile b is denoted by $\xi_i(b)$. It is easy to see that the I-tuple $\{\xi_i(b)\}$ is an attainable allocation for any $b \in B$.

The traders must return their initial balances of fiat money to the government at the end of the game. This means that trader i must end the trading period with at least m_i units of money. We assume that any choice of b_i resulting in end-of-period money balances that are lower than m_i will yield a payoff of $-\infty$. The terminal balance for

trader i equals his initial balance m_i minus the sum of his bids $\sum_{h=1}^{\ell} b_{ih}$ plus the revenue from the sale of his initial portfolio $p(b)\cdot e_i$. It is easy to show that the terminal balance satisfies

$$m_i - \sum_{h=1}^{\ell} b_{ih} + p(b)\cdot e_i = m_i - p(b)\,.\,(\xi_i(b) - e_i)\cdot$$

so the terminal constraint is satisfied if and only if $p(b)\cdot(\xi_i(b) - e_i) \leq 0$. For any strategy profile b, let trader i's payoff be denoted by $\pi_i(b)$ and defined by

$$\pi_i(b) = \begin{cases} u_i(\xi_i(b)) & \text{if } p(b)\cdot(\xi_i(b) - e_i) \leq 0, \\ -\infty & \text{if } p(b)\cdot(\xi_i(b) - e_i) > 0. \end{cases}$$

Shapley and Shubik (1977) demonstrate the existence of a Nash equilibrium for this game under the additional assumption that for each commodity h there are at least two individuals whose utility is increasing in that commodity. They also provide conditions under which the equilibrium allocation converges to a competitive equilibrium as the number of traders increases without bound.

Concluding Remarks

As Joseph Ostroy wrote in the first edition of *The New Palgrave* (1987, p. 515).

We shall argue that the incorporation of monetary exchange tests the limits of general equilibrium theory, exposing its implicitly centralized conception of trade and calling for more decentralized models of exchange.

That comment is just as true today as it was then, and remains the great challenge for economists who want to develop more satisfactory models of the process of monetary exchange at the level of the economy as a whole.

See Also

- Monetary Approach to the Balance of Payments
- Monetary Cranks
- Monetary Policy, History of
- Money Illusion
- Money Supply

Bibliography

Archibald, G., and R. Lipsey. 1958. Monetary and value theory: A critique of patinkin and lange. *Review of Economic Studies* 26: 1–22.

Balasko, Y., and D. Cass. 1989. The structure of financial equilibrium: I. exogenous yields and unrestricted participation. *Econometrica* 57: 135–162.

Cass, D. 2006. Competitive equilibrium with incomplete financial markets. *Journal of Mathematical Economics* 42: 384–405.

Clower, R. 1967. A reconsideration of the microfoundations of monetary theory. *Economic Inquiry* 6: 1–8.

Fisher, I. 1963. *The Purchasing Power of Money,* rev. edn. New York: Kelley.

Geanakoplos, J., and A. Mas-Colell. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47: 22–38.

Geanakoplos, J., and P. Dubey. 1992. The value of money in a finite-horizon economy: A role for banks. In *Economic analysis of markets and games*, ed. P. Dasgupta et al. Cambridge, MA: MIT Press.

Grandmont, J.-M. 1983. *Money and value: A reconsideration of classical and neoclassical monetary theories*. Cambridge/Paris: Cambridge University Press/Maison des Sciences de l'Homme.

Grandmont, J.-M., and G. Laroque. 1975. On money and banking. *Review of Economic Studies* 42: 207–236.

Grandmont, J.-M., and Y. Younes. 1973. On the efficiency of monetary equilibrium. *Review of Economic Studies* 40: 149–165.

Hahn, F. 1965. On some problems of proving the existence of an equilibrium in a monetary economy. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.

Hicks, J. *1946. Value and capital: An inquiry into some fundamental principles of economic theory.* 2nd ed. *Oxford*: *Clarendon*.

Keynes, J. *1936. The general theory of employment, interest and money. London*: *Macmillan*.

Lucas, R. 1980. Equilibrium in a pure currency economy. *Economic Enquiry* 18: 203–220.

Lucas, R., and N. Stokey. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12: 55–93.

Lucas, R., and N. Stokey. 1987. Money and interest in a cash-in-advance economy. *Econometrica* 55: 491–513.

Magill, M., and M. Quinzii. 1992. Real effects of money in general equilibrium. *Journal of Mathematical Economics* 21: 301–342.
Ostroy, J.M. 1987. Money and general equilibrium. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, Vol. 3. London: Macmillan.
Patinkin, D. *1965*. *Money, interest, and prices: An integration of monetary and value theory*. 2nd ed. *New York*: *Harper and Row.*
Pigou, A.C. 1943. The classical stationary state. *Economic Journal* 53: 343–351.
Sargent, T. 1987. *Dynamic macroeconomic Theory.* Cambridge, MA: Harvard University Press.
Shapley, L., and M. Shubik. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85: 937–968.
Shubik, M. 1972. Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 10: 24–38.
Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36: 110–119.

Money and General Equilibrium Theory

Joseph M. Ostroy

Taking general equilibrium theory to be the model introduced by its founder, the topic of money and general equilibrium theory is as old as the subject itself. In the Preface to the fourth edition of the *Elements*, Walras wrote: 'Chiefly, however, it was my theory of money that underwent the most important changes as a result of my research on the subject from 1876 to 1899.'

We are still working on Walras's model, but while the non-monetary aspects of the model have been the subject of steady improvement marked by comparatively harmonious logical development, research on the monetary side has been greeted with doubts and misgivings. Why? There is an outward-looking reason: the subject of money in general equilibrium is thought to be related to the problems of macroeconomics, subjects of great consequence and contentiousness. There is also an inward-looking reason: it is not clear if what we know as Walrasian general equilibrium is compatible with a model in which money as a medium of exchange plays an essential role.

This essay will take the inward-looking perspective on money and general equilibrium theory. No claim is, or need be, made that only after the inward-looking issues of logical consistency have been settled will the problems of money and macroeconomics be resolved. The only claim we make is the rather obvious one that monetary exchange is an example *par excellence* of a universal economic phenomenon, and if there is one branch of the discipline that is suited to its study, it is certainly general equilibrium theory. We shall argue that the incorporation of monetary exchange tests the limits of general equilibrium theory, exposing its implicitly centralized conception of trade and calling for more decentralized models of exchange.

The Walras–Hicks–Patinkin Tradition: Integrating Monetary into Value Theory

Walras presents his framework first by addressing the problem of equilibrium in exchange and then introducing production and capital accumulation as extensions of the basic model of exchange. With this structure in place, Walras brings in money by introducing the equation of the offer and demand for money so as to conform with the rest of his system. This is accomplished by making a distinction between the stock of money, an object without any utility of its own, and the 'services of availability' of the stock, which does enter into household utility functions and firm production functions.

Similarly, Hicks's (1935) suggestion for simplifying the theory of money is to make it conform to the (non-monetary) theory of value. Since 'marginal utility analysis is nothing other than a general theory of choice' and people do choose the amounts of money they hold, monetary theory can be embedded into value theory.

Patinkin's work (1965) represents the culmination of this tradition. Here Walras's 'service of availability' of money is somewhat fancifully recast as avoidance of 'embarrassment from

default'. Important to the money and general equilibrium agenda for Patinkin is the proper formulation of the real-balance effect so that nominal and real magnitudes are jointly determined as well as more precise statements of propositions on the short-run and long-run neutrality of money.

The presumption in this integration of money into value theory is that monetary theory is the weak partner and that by the exercise of reshaping it to fit the more rigorous choice-theoretic principles of value theory, including capital theory, monetary theory will be strengthened. There can be no doubt as to the influence of this regimen. Numerous contributions have demonstrated that the mechanism of exchange in a money economy, whatever it may be, can be usefully approximated by the mechanism of choice for a money commodity. Writers such as Friedman (1956) and Tobin (1961) each subscribe to the incorporation of money into the framework of value theory as the basis for their outward-looking approach to monetary theory.

When subjected to the scrutiny of the inward-looking approach to money and general equilibrium, this goal of integration does not appear to be very satisfying. By introducing money after he had completed his theory of exchange, Walras clearly made monetary phenomena an optional add-on rather than an integral component of the mechanism of exchange. Further, it was an add-on that would have to be valued for its own sake rather than as a component enhancing the performance of the rest of the system.

A succinct illustration of the inability of the model to leave room for monetary exchange is Walras's Theorem of Equivalent Distributions. Let p, x_i, and w_i stand for prices, individual i's final allocation and i's initial distribution of commodities, respectively, all elements of a given vector space. If $[p,(x_i)]$ is an equilibrium final allocation for individuals having certain tastes and initial distributions of commodities (w_i), then $[p,(x_i)]$ is also an equilibrium final allocation for individuals having the same tastes and any other initial distributions ons (w_i') such that $\sum w_i' = \sum w_i'$ and $pw_i' = pw_i$ for all i. Thus the no-trade distribution $(w_i' = x_i)$ is in the same equivalence class with an initial distribution in which the pattern of net trades among individuals and commodities is much less trivial. Trade or no trade, it is all the same to this model of exchange.

Transactions Costs

The inward-looking approach to money and general equilibrium asserts that Walras's class of equivalent redistributions is much too coarse, certainly too coarse to provide a role for the exchange facilitating properties of money. One way to refine these equivalence classes is to revise the conventional budget-constraint, $p(x_i - w_i) = 0$, by postulating that the value of all commodities purchased, $p(\max[x_i - w_i,0])$, cannot exceed the value of the beginning of the period holdings of, say commodity 1, $p_1 w_i 1$. This is the so-called 'cash-in-advance' constraint of Clower (1967). The presumed real-world inferiority of barter exchange – purchasing commodities directly with other commodities – compared to money becomes a given. Of course, the added monetary constraint begs the question as to why it is necessary, particularly since as an added constraint it cuts down on the opportunities available under Walrasian barter. What is needed is a more comprehensive approach from which we may derive as a conclusion something resembling the cash-in-advance constraint as a solution to the problem of economizing on transactions costs.

Monetary exchange does not follow automatically once the costs of transacting are introduced. The costs of trading A for B directly must be greater than the indirect trade of A for money and then money for B. Oft-repeated lists of the properties of money (portability, durability, divisibility, etc.) call attention to attributes of an object with lower costs of exchange, but these lists merely describe the desirable features of a common medium of exchange that has already been adopted rather than provide an explanation of why the adoption should take place.

Just as in the single-period version of an exchange economy (characterized by the Theorem of Equivalent Distributions) where there is no role for money, the same conclusion holds for a multi-period extension with futures markets.

After indexing commodities by date and contingencies the model is indistinguishable from the one-period version. The key is that the individual faces a single budget-constraint for trades over the entire time horizon. Now modify this intertemporal model by making transactions costly, particularly that futures transactions are more costly than spot transactions. Thus, we leave behind models where the Theorem of Equivalent Redistributions holds, but we do not necessarily enter the world of monetary exchange. In fact, if we again permit individuals to face a single budget-constraint, a pattern of exchange that could be identified as monetary would require that one commodity is singled out to have much lower costs of transacting whenever it is used to buy or sell any other commodity. The question then would be 'Why?' and the answer 'That this is simply a feature of the transactions technology' would not be very satisfactory.

Suppose, however, we use the time index to create breaks in the budget-constraint. Each individual faces a sequence of budget-constraints in each of which his/her trades must balance. This will typically lead to a Walrasian equilibrium allocation that is Pareto-inferior. (Note that the definition of Pareto-inferiority takes transactions costs for granted, i.e., feasible reallocations must respect the given transactions technology and distribution of initial endowments.) Now introduce an additional object of exchange, money, the terminal holding of which must coincide with its initial holding for each individual. Then it is possible to have budget-balance in each period in all commodities including money without having period-by-period balance in nonmoney commodities. The end result is to return to the single budget-constraint for the nonmoney commodities over the whole trading horizon. Hahn (1973) and Starrett (1973) show that an equilibrium allocation under such an arrangement would be Pareto-optimal.

The moral we draw from this story is that there are two types of transactions costs, technological and strategic. The technological ones have a transportation cost character – the unavoidable costs of sending commodity A from person i at time t to person j at time s. They may set the stage, but they are not sufficient to rationalize monetary exchange. The strategic costs are reflected by the demand that budget-balance be imposed at each period. Presumably, if individuals were not required to balance their budgets at each period there would be no monitoring and enforcement mechanism to get them to balance their budget over time, and they would cheat. It is the implicit costs of monitoring and enforcing budget-balance – the strategic costs – that yield a rationale for monetary exchange.

Money and the Overlapping Generations Models of General Equilibrium

It is useful to think of general equilibrium models as coming in two versions, the predominant one due to Walras and another, called the overlapping generations model, due to Samuelson (1958). Both share what seem to be the main features of market clearance conditions obtained from price-taking maximizers, but in one important conclusion they diverge. With inessential qualifications, a Walrasian equilibrium is always Pareto-optimal while the corresponding pricetaking, market-clearing equilibrium in an overlapping generations model readily admits the possibility of Pareto inefficiency. The presence of this Pareto-gap holds out the promise that it might be filled by the introduction of a tradable asset which, although intrinsically worthless, would allow individuals of adjacent generations to reach a Pareto-optimal allocation. And this promise can be fulfilled.

There is a certain similarity between the rationale for money in the transactions costs and overlapping generations models. In each case an intertemporal equilibrium without money may be inefficient, while the introduction of an intrinsically worthless object of exchange can remove the inefficiency. Ignoring the subtle mathematical complexities of the overlapping generations model's double infinity of individuals and commodities, the hypothesis that time and future generations are unending can be accepted as a fact-of-life. Thus, without having to appeal to transactions costs, it has been boldly argued by Wallace

(1980) and Cass and Shell (1980) that the overlapping generations framework is the natural vehicle for describing money in general equilibrium theory.

Taking an outward-looking view of the problem, there is much to recommend the overlapping generations models. They lead to definite, policy-relevant macroeconomic conclusions, whereas the transactions cost approach has, at this stage at least, little to say about policy. However, taking an inward-looking view, the overlapping generations model appears less satisfying. Certainly it provides a role for money as a transfer mechanism between generations but there is no role for money as a medium of exchange. There are many circumstances in which full Pareto-efficiency can be achieved in a non-monetary equilibrium and the conditions under which efficiency does or does not require a positively valued money follows a logic of its own independent of the exchange enhancing properties commonly associated with money.

Money and Decentralized Exchange

In comparison to the aggregative style of macroeconomics, general equilibrium theory is held out as the micro-economically detailed description of an economy which highlights the decentralized character of the price system. In the Walras–Hicks–Patinkin tradition, general equilibrium theory provides the standard of rigour and detail to which monetary theory should aspire. However, when one adopts an inward-looking view of the problem of money and general equilibrium, it becomes apparent that these aspirations are set too low. The (implicit) description of market exchange in general equilibrium theory exhibits a substantial amount of as-if centralization, certainly too much to permit a role for money. Alternatively put, the Walrasian model of exchange is not much concerned with how commodities are exchanged.

Suppose the mythical auctioneer has just completed the task of finding equilibrium prices. It now remains for trades to be executed. Consider, first, a centralized story in which individuals come to the auctioneer to make their exchanges. Assuming that the auctioneer has no inventories of commodities and that not everyone can converge on the auctioneer at once, a record-keeping problem emerges. All individuals will leave their excess supplies with the auctioneer but at least the first few will not be able to pick up all their excess demands. They will have to return at a later date. Thus, actual purchase and sale will be separated in time. It would, therefore, be advisable for the auctioneer to keep a record of each individual's transactions. This can be simply and conveniently accomplished by issuing an IOU recording the value of supplies given up minus purchases received, all computed at equilibrium prices. In this way the auctioneer does not have to rely on his memory to discourage those who would cheat by saying that they had given more or taken less than they actually had in their previous trips to the auctioneer.

The strategic issues are similar to those described in the transactions cost models, above, except that here the record-keeping problem occurs whenever there is trade and not simply when there is intertemporal trade in the general equilibrium sense. Physical limitations on the executions of trades make it either inefficient or impossible to balance purchase sale transactions at every trading opportunity. But this creates the problem of how to enforce the overall budget-constraint when efficient execution of trades requires that an individual's trading position be out of balance along the way.

The auctioneer story is rather centralized. We may also consider a more decentralized trading arrangement in which individuals exchange sequentially in pairs. Ostroy (1973) and Ostroy and Starr (1974) investigate the trade-offs among time, information required beyond knowledge of equilibrium prices, inventories of commodities on hand, and centralized enforcement of budget constraints to execute trades. Also using a model of sequential pairwise exchange, Jones (1976) has addressed a theme which goes to the heart of the decentralization issue. The issue, raised by Menger (1871), is whether money is necessarily a creature of the state or whether a monetary trading pattern could arise endogenously through individuals being led by their self-interest to

single out some commodity as a common medium of exchange.

There is hardly a more universal economic phenomenon than monetary trade. Thus, it would seem an explanation of money would be at the core of a theory of exchange. That it is not is neither a cause of anguish nor for complacency. The received theory arose to explain the prices of commodities. While obviously well-suited to its purpose, it is simply too centralized to cope with the economic issues underlying monetary exchange. But current research indicates that economic theory is moving along several fronts towards a more decentralized level of abstraction. Complementary developments in theories of search, of contracting and of incentive compatibility are all examples of what is sometimes called a 'micro-micro' attempt to go beyond the levels of aggregation that constitute the more traditional modes of analysis in general equilibrium theory. Perhaps in several decades we shall look back on traditional general equilibrium theory and say that in its microeconomic detail it stands in relation to the new theory as classical Ricardian analysis stands in relation to it. At that point, we can expect monetary exchange to be a routine application of general equilibrium theory.

See Also

- ▶ General Equilibrium
- ▶ Monetary Disequilibrium and Market Clearing
- ▶ Money and General Equilibrium Theory
- ▶ Money, Classical Theory of
- ▶ Neutrality of Money
- ▶ Overlapping Generations Model of General Equilibrium
- ▶ Quantity Theory of Money
- ▶ Uncertainty
- ▶ Walras's Law

Bibliography

Cass, D., and K. Shell. 1980. In defense of a basic approach. In *Models of monetary economics*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6(December): 1–8.

Friedman, M. 1956. The quantity theory of money – A restatement. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Hahn, F.H. 1973. On transactions costs, inessential sequence economics and money. *Review of Economic Studies* 40(4): 449–461.

Hicks, J.R. 1935. A suggestion for simplifying the theory of money. *Economica* 2(February): 1–19.

Jones, R.A. 1976. The origin and development of media of exchange. *Journal of Political Economy* 84(4), Part I: 757–775.

Menger, C. 1871. *Principles of Economics*. Trans. J. Dingwell and B. Hoselitz. New York: New York University Press.

Ostroy, J.M. 1973. The informational efficiency of monetary exchange. *American Economic Review* 63(4): 597–610.

Ostroy, J.M., and R. Starr. 1974. Money and the decentralization of exchange. *Econometrica* 42(6): 1093–1113.

Patinkin, D. 1965. *Money, interest and prices*, 2nd ed. New York: Harper & Row.

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.

Starrett, D. 1973. Inefficiency and the demand for 'money' in a sequence economy. *Review of Economic Studies* 40(October): 437–448.

Tobin, J. 1961. Money, capital and other stores of value. *American Economic Review: Papers and Proceedings* 51: 26–37.

Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of monetary economics*, ed. J.H. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.

Walras, L. 1874–7. *Elements of Pure Economics*. Trans. W. Jaffé. Homewood: Richard D. Irwin, 1954.

Money Illusion

Peter Howitt

JEL Classifications
C0

The term money illusion is commonly used to describe any failure to distinguish monetary from real magnitudes. It seems to have been

coined by Irving Fisher, who defined it as 'failure to perceive that the dollar, or any other unit of money, expands or shrinks in value' (1928, p. 4). To Fisher, money illusion was an important factor in business-cycle fluctuations. Rising prices during the upswing would stimulate investment demand and induce business firms to increase their borrowing, thus causing a rise in the nominal rate of interest. Lenders would accommodate them by increasing their savings in response to the rise in the nominal rate, not taking into account that, because of the rise in inflation, the real rate of interest had not risen but had actually fallen (Fisher 1922, esp. ch. 4).

Beginning with Haberler (1941, p. 460, fn. 1) other writers have used the term money illusion as synonomous with a violation of what Leontief (1936) called the 'homogeneity postulate', the postulate that demand and supply functions be homogeneous of degree zero in all nominal prices; that is, that they depend upon relative prices but not upon the absolute price level. This usage differs from Fisher's in two senses. It refers to people's reactions to a change in the level of prices rather than to a change in the rate of change of prices, and it is cast in operational terms, as a property of potentially observable supply and demand functions rather than as a property of people's perceptions or lack thereof.

Patinkin (1949) objected to the latter use of the term money illusion on the grounds that it failed to take into account the real balance effect. A doubling of all money prices should affect household demand functions even if people are perfectly rational and suffer from no illusions, because it reduces at least one component of the real wealth that constrains their demands – the real value of their initial money holdings. Accordingly he defined the absence of money illusion as the zero-degree homogeneity of net demand functions in all money prices and the money values of initial holdings of assets.

In a fiat money economy in Hicksian temporary equilibrium, under the assumption of static expectations, the absence of money illusion in Patinkin's sense is operationally equivalent to the assumption of rational behaviour, in the following sense. Let each agent's demand functions $x_i^{\wedge}(p_1,\ldots,p_n, W)$ for goods $i = 1, \ldots, n$, together with his demand-for-money function $M(p_1,\ldots, p_n, W)$ be defined as the maximizers of the utility function $U(x_1,\ldots, x_n; M,\ldots, p_n)$ subject to the budget constraint: $p_1x_1 + \ldots + p_nx_n + M = W$, where W is initial nominal wealth. The utility function includes M and the money prices p_i because M is assumed to yield unspecified services whose value depends upon the vector of prices expected to prevail next period, and those expected prices are proportional to today's prices.

A rational agent would realize that a proportional change in M and all prices would leave unaffected the purchasing power of M, and thus also the services rendered by M. Accordingly U is said to be illusion-free if it is homogeneous of degree zero in $(M, p_1, \ldots, p_n)$. This homogeneity property was first assumed explicitly in the context of demand theory by Samuelson (1947, p. 118) although it was implicit in the earlier analysis of Leser (1943), who used the equivalent formulation: $U(x_1, \ldots, x_n; M/p_1, \ldots, M/P_n)$. It is easily verified that the $\widehat{x}$'s are illusion-free in Patinkin's sense if and only if they can be derived from an illusion-free U (see Howitt and Patinkin 1980).

The assumption of static expectations is crucial to this equivalence. If expected future prices were not proportional to current prices then a proportional change in $p_1, \ldots, p_n, W$ would alter intertemporal relative prices and it would not be irrational for the agent to respond by changing his demands. Patinkin's original definition can be generalized to take this possibility into account and to allow for the presence of productive non-money assets by requiring demand functions for real goods to be unaffected by a proportional change in W, all current prices, and all expected future prices, holding constant the rates of return on all non-money assets. If future prices p_i' were uncertain then current demands would depend upon the probability distribution $F(p_i',\ldots,p_n')$, and the proportional change in future expected prices in the above statement would have to be replaced by a change from $F(p_i', \ldots, p_n')$ to $F_\lambda(p_i',\ldots,p_n') \equiv F(p_l'/\lambda,\ldots,p_n'/\lambda)$ where λ is the factor of proportionality.

The absence of money illusion is the main assumption underlying the long-run neutrality proposition of the quantity theory of money. But the presence of money illusion has also frequently been invoked to account for the short-run non-neutrality of money, sometimes by quantity theorists themselves, as in the case of Fisher. On the other hand, many monetary economists have reacted adversely to explanations based on such illusions, partly because illusions contradict the maximizing paradigm of microeconomic theory and partly because invoking money illusion is often too simplistic an explanation of phenomena that do not fit well into the standard equilibrium mould of economics. Behaviour that seems irrational in a general equilibrium framework may actually be a rational response to systemic coordination problems that are assumed away in that framework.

Thus, for example, Leontief (1936) attributed Keynes's denial of the quantity theory to an assumption of money illusion. He interpreted Keynes as saying that the supply of labour depended upon the nominal wage rate whereas the demand depended upon the real wage. A rise in the price level would thus raise the equilibrium quantity of employment. Leijonhufvud (1968, ch. 2) questioned this interpretation and argued that Keynes was dealing with information problems that don't exist in Leontief's general equilibrium analysis. Specifically, Leijonhufvud argued that workers might continue supplying the same amount of labour services in the event of a rise in the general price level, not because they irrationally identified nominal with real wages but because in a world of less than perfect information it would take time for them to learn of the changed value of money.

Likewise, Friedman (1968) objected to the then standard formulation of the Phillips-relation between unemployment and the rate of wage-inflation. Friedman argued that the rate at which firms raised their wage offers and households raised their reservations wages, given any existing amount of unemployment, should depend upon these agents' expectations of the future value of money. To assume otherwise would be to assume money illusion. Friedman's argument implied that an expected-inflation term should be added to the usual specification of the Phillips curve. His analysis of the expectations-augmented Phillips curve was similar to Leijonhufvud's imperfect-information argument.

More recently, Barro (1977) has argued against the assumption of nominal wage stickiness in the work of Fischer (1977) and others, on the grounds that microeconomic theories of wage contracts imply that these contracts should be signed in real, not nominal terms, unless people suffer from money illusion.

Although monetary economists have thus been reluctant to attribute money illusion to private agents they have not hesitated to attribute it to governments. Indeed, as Patinkin (1961) demonstrated, money illusion on the part of the monetary authority is necessary for an economy to possess a determinate equilibrium price level. More recently, several writers have attributed real effects of inflation to money-illusion in tax laws (e.g., Feldstein 1983). Specifically, in many countries interest income and expenses are taxed at the same rate regardless of the rate of inflation, and historical money costs rather than current replacement costs are used for evaluating inventories and calculating depreciation allowances. Because of these effects inflation can distort the after-tax cost of capital.

In short, the attitude of economists to the assumption of money illusion can best be described as equivocal. The assumption is frequently invoked and frequently resisted. The persistence of a concept so alien to economists' pervasive belief in rationality indicates a deeper failure to understand the importance of money and of nominal magnitudes in economic life. This failure is evident, for example, in the lack of any convincing explanation for why people persist in signing non-indexed debt contracts, or why the objective of reducing the rate of inflation, even at the cost of a major recession, should have such wide popular support in times of high inflation.

See Also

► Neutrality of Money
► Real Balances

Bibliography

Barro, R.J. 1977. Long-term contracting, sticky prices, and monetary policy. *Journal of Monetary Economics* 3 (3): 305–316.
Feldstein, M. 1983. *Inflation, tax rules, and capital formation*. Chicago: University of Chicago Press.
Fischer, S. 1977. Long-term contracts, rational expectations, and the optimal money supply rule. *Journal of Political Economy* 85 (1): 191–205.
Fisher, I. 1922. *The purchasing power of money*. 2nd ed. New York: Macmillan.
Fisher, I. 1928. *The money illusion*. New York: Adelphi.
Friedman, M. 1968. The role of monetary policy. *American economic review* 58 (1): 1–17.
Haberler, G. 1941. *Prosperity and depression*. 3rd ed. Geneva: League of Nations.
Howitt, P., and D. Patinkin. 1980. Utility function transformations and money illusion: comment. *American Economic Review* 70 (3): 819–822.
Leijonhufvud, A. 1968. *On Keynesian economics and the economics of Keynes*. New York: Oxford University Press.
Leontief, W. 1936. The fundamental assumptions of Mr Keynes' monetary theory of unemployment. *Quarterly Journal of Economics* 5 (November): 192–197.
Leser, C.E.V. 1943. The consumer's demand for money. *Econometrica* 11 (2): 123–140.
Patinkin, D. 1949. The indeterminacy of absolute prices in classical economic theory. *Econometrica* 17 (1): 1–27.
Patinkin, D. 1961. Financial intermediaries and the logical structure of monetary theory. *American Economic Review* 51 (1): 95–116.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Money in Economic Activity

D. Foley

Money is a social relation. Like the meaning of a word, or the proper form of a ritual, it exists as part of a system of behaviour shared by a group of people. Though it is the joint creation of a whole society, money is external to any particular individual, a reality as unyielding to an individual's will as any natural phenomenon.

To understand the system of social relations of which money is a part, it is necessary to adopt a comparative and historical perspective. Only by seeing the phenomenon of money in contrast with systems of social relations that do not involve money can we get a sense of the characteristic peculiarities of money. Marx's (1867) analysis of commodity production provides such a perspective.

People in every society must produce in order to survive and develop, but their production can be organized through different systems of social relations. An important dimension of these social relations is the degree to which the products are controlled by individual owners acting in their own interests. In a system of commodity production, a product at its creation is the property of one owner, who can exchange it for the products owned by others.

Money appears in systems of commodity production. Because any commodity can be transformed into any other through exchange, commodities appear to be equivalents. It is possible to evaluate any collection of disparate commodities by multiplying the quantity of each one by a price, where the ratio of the prices of any two commodities expresses the ratio in which they exchange, and adding up. Because exchange determines only the ratio of the prices, the units in which value is measured are arbitrary. A similar situation exists in measuring the mass of physical objects. By weighing one mass against another one can establish the proportion of one to another, but to express weight in absolute terms it is necessary to establish a conventional standard (like the kilogram or pound). In a commodity-producing society some system evolves for measuring and transferring value separated from particular commodities, the money form of value. Monetary units such as the dollar, franc, pound, mark, or yen, measure value separated from particular commodities.

Although the money form of value is a universal characteristic of commodity systems of production, different specific forms of money have evolved in different times and places. The earliest form of money is commodity money. One particular product, often a precious metal such as gold, takes on the role of measuring the value of all other commodities. In a commodity money system the monetary unit, for example the dollar, is

defined legally as a certain amount of gold. Since gold exchanges at a particular ratio with every other commodity, this definition establishes a dollar price for every commodity as well.

It is also possible for commodity systems to operate with an abstract unit of value, a monetary unit implicitly defined by prices negotiated by buyers and sellers of commodities. In this situation, the dollar is not defined as a particular quantity of some commodity, but commodity producers, knowing at any moment how much value the dollar represents, continue to establish prices in terms of dollars. Commodity money systems are subject to instability because the exchange ratios of the money commodity against other commodities constantly change. Abstract unit of account systems are subject to instability because the prices producers choose may drift upward or downward over time.

In a commodity money system agents may issue promises to pay a certain amount of money at a particular time in the future, or on demand. These promises to pay, liabilities for the issuer and assets for the holder, if they are credible, can take the place of the money commodity in many situations. For example, if a producer agrees to sell his product for a certain money price, he may accept a credible promise to pay gold instead of gold itself. Likewise, if an individual needs to hold a stock of money to provide for contingencies, she may decide to hold widely acceptable promises to pay rather than gold itself, if that is more convenient. The same thing can happen in an abstract unit of account system. Promises to pay pure value may be acceptable in transactions, and used as stores of value.

In systems where credit is highly developed, what does it mean for one agent to promise to pay money? How can this promise be fulfilled? In a commodity money system, payment ultimately means delivery of a certain quantity of the money commodity. In an abstract unit of account system, payment normally means delivering a third party's promise to pay, where the third party's liability is more acceptable than the debtor's. For example, private producers may pay each other by transferring the liabilities of banks, deposits. Banks in turn may pay each other by transferring the liabilities of the State, bank reserves or currency. In a commodity money system every agent faces an ultimate requirement to pay in the money commodity. In an abstract unit of account system, however, the State does not have to pay its liabilities by transferring something else.

It is surprising how little difference there is in the day-to-day practice of systems with and without a money commodity. For most individual agents there is one type of highly acceptable liability (bank deposits, for instance) in which the agent must settle its accounts. The same thing is true in a money commodity system. The fact that at the top of the pyramid of agents whose liabilities are more and more socially acceptable the State has to pay in gold in a commodity money system, and does not have to pay any particular thing in an abstract unit of account system makes no difference to the individual agents. Even in a commodity money system, the development of a pyramid of agents whose liabilities have different degrees of acceptability insulates most of the agents in the system from the money commodity itself. Only in periods of crisis, when the State faces severe difficulties in maintaining the convertibility of its liabilities the money commodity, will the money commodity influence the financial decisions of individual agents.

Liabilities of high social acceptability, like currency issued by the State, or bank deposits, may come to be preferred as the means of payment for individual transactions, though in almost all commodity producing societies other liabilities also perform important payment functions. For example, endorsed bills of exchange of private traders have often circulated as widely accepted means of payment among firms in capitalist societies. Furthermore, the issuers of liabilities of high acceptability find that agents are willing to hold them even when they pay a lower rate of return than other assets. If the issuers of these liabilities can exercise some monopoly power, as banks organized under the leadership of a State-sponsored central bank can, they will restrict the interest paid on their liabilities to a minimum. This minimum may in some cases reach zero, so that the most socially accepted liabilities pay a zero interest

rate. Agents continue to hold these liabilities as their assets because of their wide acceptability as payment, and because they serve very well as a reserve against the contingency that the agent will not be able to borrow.

From this examination of the nature of money as a social relation, we can draw several important conclusions on which to base a discussion of money and economic activity. First, the money form of value, value separated from a particular commodity, is inherent in the organization of production through exchange. Second, the emergence of money takes place simultaneously with the growth of exchange itself. Third, while the money form of value is a universal characteristic of commodity production, the forms of money are diverse and changing. In particular, the liabilities, or promises to pay, of economic agents can serve in place of a money commodity, and can take the place of the money commodity altogether. Fourth, whether there is or is not a money commodity, there tends to develop a hierarchy of liabilities of different degrees of acceptability. Payment for agents at one level of this pyramid requires their delivery of liabilities of agents at the next level up. The existence of this pyramid creates considerable flexibility in the financing of economic activity.

The relation between money and economic activity is two-sided. Money forms of value are a reflection of the particular organization of economic activity through commodity production. The liabilities that serve to finance economic activity are created in the course of financing economic activity itself. From this perspective it is tempting to argue that money is a reflector of economic activity, and that monetary phenomena are determined by the independent development of economic activity. As we shall see, this is an important theme in the development of monetary theory.

But money also serves as a regulator of economic activity, because it is the link between the individual producer and the social character of production. In order to undertake production, an agent must finance it by getting control of an acceptable monetary asset. If an agent does not already own a sufficient quantity of these assets, it must convert its own liability into a liability of higher acceptability by borrowing. The terms on which agents can make this transformation regulate their initiation of production in two senses. First, financing determines which agents will be able to carry out their plans. Second, financing determines the total volume of economic activity that can be initiated. In its role as regulator of economic activity, money appears, especially from the perspective of the individual economic agent, to be the independent factor to which economic activity has to adapt itself.

Theories of money can be seen first in terms of which of these aspects of the relation between money and economic activity they emphasize as their starting point, and second in terms of the way they account for the final synthesis of the two points of view. Those theories that posit an independent role for money in determining economic activity have some level at which money is itself determined by economic activity, and those theories that emphasize money as a reflector of economic activity also envision circumstances where money regulates and influences the scale of economic activity.

In the 18th and early 19th centuries, the writers who had the most influence on the later development of monetary theory, Hume, Smith, Ricardo, and Marx, all place the main emphasis on money as a reflector of levels of economic activity determined by non-monetary factors.

David Hume (1752) makes two, somewhat contradictory, arguments concerning the reasons why the quantity of money has no lasting effect on the levels of economic activity. The first is that the money prices of commodities are proportional to the quantity of money in a country. As a result, the real quantity of money, correcting the quantity of money for the level of money prices, is endogenous. Since the real quantity of money is relevant for economic decision making, and particularly for decisions regarding the initiation of economic activity, once prices have adjusted, the physical quantity of money commodity in the country makes no difference. But in a second essay Hume argues that in fact the physical quantity of money in a country is also endogenous, here implicitly assuming that the gold prices of

commodities are determined by non-monetary factors, essentially by world prices. Here his argument is that a country with a relatively small quantity of money commodity will have low prices relative to the rest of the world, which will create a balance of trade surplus and attract the money commodity to that country. This process will continue until the price level in the country has risen to the level of world gold prices. There are two processes of adjustment in Hume's argument, a middle run in which money prices of commodities are proportioned to the quantity of the money commodity, and a long run in which, because prices of commodities are determined by world prices, the quantity of the money commodity in the country adjusts.

But Hume makes yet a third remark about the relation of money to economic activity, which raises an important theme for later writers. He argues that there is a short run in which increases in the quantity of money in a country do directly increase the level of economic activity because commodity prices have not fully adjusted to the quantity of the money commodity. Later writers attempt to flesh out this argument by specifying the exact mechanism through which changes in the nominal quantity of money can affect the level of economic activity.

Adam Smith (1776) emphasizes quite a different aspect of the relations of money to economic activity. Smith's discussion of credit and banking centres on the idea that the substitution of credit, particularly banknotes, for precious metals as a medium of circulation can free social capital tied up in stocks of the money commodity to set production in motion. In this perspective credit has a significant effect on the level of economic activity. Smith is concerned to enunciate rules of banking that will prevent an overissue of banknotes and maintain convertibility of banknotes into the money commodity, rules which are the origin of the real bills doctrine. Smith recommends that banks lend only to real creditors who are already owed money by real debtors as the result of bona fide commodity transactions. Such loans will be automatically liquidated when the real debtor pays real money (that is, the money commodity) to the creditor and the creditor in turn pays the money into the bank. Essentially Smith argues for a system in which borrowers are forced at frequent, periodic intervals to clear their positions and demonstrate their continued solvency. He claims that a banking system that follows these rules will have no difficulty in maintaining convertibility, so that its banknotes will circulate at par against the money commodity, and can replace a certain proportion of the money commodity as a medium of circulation.

Smith views a properly regulated banking system as providing the appropriate amount of money endogenously through the expansion and contraction of credit. There are two levels to Smith's argument. At the first level, the introduction of banks and credit money have a once and for all effect on economic activity by releasing social capital previously tied up in stocks of the money commodity for production. Once the banking system is in place and functioning to its maximal feasible extent according to the rules of the real bills doctrine, however, the quantity of money and credit, now endogenous to the system, has no independent effect on the level of economic activity (nor, apparently, on prices, which Smith sees as being regulated rapidly by world prices).

Both Smith and Hume are at pains to establish that the quantity of money does not influence the level of interest rates, which they view as being determined by the level of profit rates in a country. In their view there is a conventional relation between the level of profit rates and interest rates. A low interest rate reflects a low profit rate as a result of the healthy development of commodity production in a country and the exploitation of profit opportunities, not an abundance of the money commodity.

David Ricardo's (1811) thinking on monetary matters arose from his study of the problem of the price of gold bullion in terms of pounds during the Napoleonic Wars, when the Bank of England suspended the convertibility of its banknotes into gold. During this period the market price of gold bullion rose to a substantial premium over the official, mint price of gold. This prompted a debate over the reasons for the premium and the appropriate policy to deal with it. A number of people argued that the premium reflected real

factors, such as poor harvests, that had created a balance of trade deficit for England, and had driven the pound to a discount against foreign currencies defined in terms of gold. Ricardo insisted, instead, that the premium reflected an overissue of banknotes by the Bank of England. He claimed that this overissue put more notes in the hands of the public than it wanted to hold, and that in attempting to get rid of the excess, the public tried to buy gold bullion and drove up its price. For Ricardo, the policy appropriate to the situation was one of restricting the issue of banknotes as a prelude to resuming conversion of notes into gold. He further argued that any impact of real factors, like bad harvests, on the price of gold bullion must take place by way of monetary changes. In other words, in the absence of an overissue of paper currency, and a consequent rise in the price of bullion, a bad harvest would lead to a rise in other commodity exports to pay for the import of grain, not to a depreciation of the pound in terms of gold.

Ricardo's discussion raises a new question, which has great importance for the later development of monetary systems. This is the question of the effect of the issue of banknotes that, unlike Smith's convertible notes, are not convertible into the money commodity at a guaranteed rate of exchange. The broad thrust of Ricardo's argument is that the issue of such notes has no effects on the economy, because overissue simply leads to a discount of the notes against the money commodity. Once again, the quantity of real money has become endogenous, now not through changes in the prices of commodities in terms of the money commodity, but through changes in the discount of paper banknotes against the money commodity.

Ricardo later goes considerably further than this analysis and explicitly argues for the independence of levels and directions of economic activity from monetary factors. Because he believed that the only rational end of economic activity was consumption, Ricardo argued, following Say, that every commodity offered for sale represented a demand for some other commodity, and thus, that in the aggregate the value of commodities offered on the market equalled the demand. Thus money is purely a medium for the exchange of commodities against each other, and has no independent role in determining economic activity; money is a veil.

Karl Marx (1867) develops his theory of money as a critique and correction of the ideas of these earlier writers. He has three important themes of correction in his approach to money. First, he argues that the prices of commodities in a commodity money system are prior to the quantity of money, so that the quantity of money theory of the price level is mistaken. Second, he rejects Ricardo's espousal of Say's Law on the ground that the movement of money into and out of hoards may create a discrepancy between the aggregate supply of commodities and the aggregate demand. Third, Marx argues for viewing the quantity of the money commodity as endogenous to the economic system, and insists that a sharp distinction be made between the effects of exogenous issues of nonconvertible paper money, and the endogenous movements of the money commodity. Still, Marx's overall view emphasizes the primacy of production decisions limited by the accumulation of capital in regulating the level of economic activity, and portrays monetary events as primarily reflecting or communicating forces set in motion at the level of production.

In Marx's theory the money price of a commodity reflects the relation between the cost of production of the commodity and the cost of production of the money commodity. In the simplest case in which costs of production are proportional to labour times expended, this implies that the money price of a commodity is just the ratio of the labour time expended in producing it to the labour time expended in producing a unit of the money commodity. If, for example, it takes one hour of labour time to produce a bushel of wheat, and two hours to produce an ounce of gold, the gold price of a bushel of wheat will be ½ ounce of gold. In Marx's analysis monetary units, like the dollar or pound or franc, are simply conventional names for specific quantities of gold. If an ounce of gold is defined to be equal to 20 dollars, for instance, then the price of a bushel of wheat will be 10 dollars in the example above. In this way, the money commodity takes on the special role of expressing the abstract labour contained in other commodities. But this role depends on the cost of

production of the money commodity, not on the quantity of it that happens to be in a certain country at a certain time.

How, then, does the quantity of money adjust to changes in the scale of economic activity? Marx introduces the idea that agents hoard money, so that there are reserves of the money commodity available to be brought into circulation in response to increases in economic activity, and ready to absorb excess quantities of the money commodity if economic activity slackens. Marx's recognition of the existence of hoards is a key distinction between his vision of monetary theory and that of Hume and Ricardo. It leads logically to another important difference in Marx's treatment of Say's Law. Because Marx included the possibility of hoarding in his theory, he saw the possibility that the proceeds from sales of commodities might be hoarded rather than spent, thus breaking the close connection between the aggregate supply of commodities and aggregate money demand asserted by Ricardo and Say.

In his discussion of inconvertible paper money issued by the State in a system based on a commodity money, Marx returns to a position very similar to Ricardo's early analysis of the price of gold bullion. Following Smith, Marx argues that the issue of paper can displace gold without a depreciation of the paper, as long as the quantity of paper issued is smaller than the requirements of circulation. Under these circumstances all the paper will be absorbed by circulation, displacing an equal value of gold, and will circulate at par against gold. If, however, the State issues more paper than this, agents trying to dispose of the excess will bid the paper to a discount against gold. The quantity of money theory of prices holds for inconvertible paper money in Marx's view, but only through the mechanism of the premium for gold against the paper money. The quantity of gold itself has no impact on gold prices, because these are determined by costs of production.

In Marx's view the level of economic activity is regulated primarily by the historical accumulation of value as capital. At any moment the technology in use establishes capital requirements for the production of various commodities. The amount of capital value available from past accumulation sets a limit to the scale of economic activity. Money in normal circumstances adapts passively to this level, either through the adjustment of hoards, or through the expansion and contraction of credit. In periods of crisis, however, the stagnation of money in reserve hoards is for Marx the mechanism by which aggregate demand is reduced. Marx's account of the exact relation of economic activity to money in periods of crisis is incomplete. It is clear that he viewed the existence of money, and the possibility of hoarding as preconditions for crisis, and as important channels in the development of crises. He also strongly suggests that the underlying causes of crises lie in the evolution of production itself, for example, in the tendency of rate of profit to fall with capital accumulation and capitalist development of production.

The classical economists and Marx left a well-developed account of the relation of money to economic activity, an account which shaped later thinking in decisive ways. These theorists assumed unquestioningly that they were dealing with a commodity money system. The only exception to this rule is the analysis of inconvertible paper money issued by the State, and coexisting with a commodity money system. The characteristic theme of classical analysis was the subsidiary importance of money in relation to production. Money was seen as adapting to economic activity, either by automatic adjustments in the quantity of money, or in real quantities of money through changes in the prices of commodities.

The century after 1875 was a period of rapid and thoroughgoing transformation of monetary systems and financial institutions in the industrialized capitalist countries. With the growth of national markets and firms operating on a national and, increasingly, international scale, national markets for credit also developed. Large banks began to play a central role in the mobilization and channelling of national capital funds. Periodic monetary panics, involving external or internal drains of gold from the reserves of banks, began to occur. National banking systems became oligopolized and regulated. Thus the monetary

phenomena that Smith, for example, analysed in the context of a largely agrarian and commercial capitalist society came to play a decisive role in the financing and construction of large-scale industrial development.

The important capitalist nations during this period extended their influence over the whole rest of the world in the first wave of capitalist imperialism. The world monetary system came to play a more and more important part in regulating economic activity on a world scale. The rivalries intensified by imperial competition between European powers set off a chain of disastrous social and military crises, beginning with World War I. The world monetary system was fundamentally and irreversibly transformed by these crises. During the war, all the participant nations suspended convertibility of their currencies into gold, and erected elaborate systems of control over movements of capital. As a result the link between gold and national currencies became much weaker. The governments of the European powers discovered that their domestic monetary and credit mechanisms depended very little on convertibility for their day to day functioning. They also discovered the enormous latitude for State policies opened up by their abandonment of the promise to convert currency into gold. Although most political leaders expected the gold standard to return after the war, commitments to convertibility turned out to be fragile and temporary. Since 1914, convertibility of national currencies into a commodity money has been the exception rather than the rule, attainable only for short periods as the result of intensive diplomatic compromise.

The earlier monetary theory we have discussed might lead one to predict that under these circumstances national currencies would gradually lose their monetary role in competition with a spontaneously maintained world commodity money standard, so that all the national currencies would find their own discount or premium against gold, which would still function as a commodity money. While something like this did occur between the First and Second World Wars, after World War II a surprising evolutionary development occurred, in which one national currency, the dollar, despite its tenuous and tentative convertibility into gold, emerged as a world monetary standard. When the United States finally abandoned convertibility of the dollar into gold in 1971, it became clear that gold had lost its central position in the world monetary system. The industrialized world functioned with the dollar, an abstract unit of account, whose value in terms of commodities is determined by the pricing decisions of commodity buyers and sellers, as the standard of value.

These historical and institutional developments called into question much of classical monetary theory, which was based on the assumption of a commodity money system. In particular, those theories that argued that the value of the monetary standard was determined by the cost of production of the money commodity were left with the need to propose an alternative mechanism for determining the value of the monetary unit. The development of monetary theory in this period reflects the attempts of economists to grapple with this fundamental problem.

Irving Fisher (1911), writing in the heyday of US trustification about the turn of the 20th century, returned to the simplest formulation of the quantity of money theory of prices put forward by Hume as the starting point for his account of the relation between money and economic activity. Fisher posits the existence of a given amount of money, exogenously determined in the system. Because he assumes that this total quantity of money must circulate (thereby abstracting from the possibility of hoards) at a single exogenously given rate (the velocity of money), Fisher argues that the total monetary value of the transactions in an economy is determined independently of the level of economic activity. If, for example, there exist $100 billion dollars, exogenously supplied, and the velocity of money is five transactions per year on average, then the total transactions of the economy must total $500 billion per year. How can this be reconciled with the actual level of economic activity? Either the volume of transactions at given prices must change so that the total equals $500 billion, or the prices at which transactions occur must change to achieve the same result. Fisher followed Hume in arguing that,

while in the short run a change in the quantity of money or velocity might have some impact on the level of economic activity in the society, in the long run the whole adjustment would be made in the prices of commodities. Fisher believed that the market system would lead to a given level of production of commodities determined by available resources and technological possibilities independently from monetary factors. Thus the only remaining variable free to adjust is the level of commodity prices. Fisher resurrects the classical presupposition that monetary factors do not influence economic activity, at least in the long run, on the basis of this analysis.

The monetary theory of John Maynard Keynes (1936) responds to the drastic changes in monetary systems engendered by World War I and the Great Depression. Keynes envisions a monetary system with a central bank at its centre. The liabilities of this bank may or may not be convertible into a money commodity. The liabilities of the central bank serve as the reserves of a commercial banking system which issues deposits. Keynes explicitly allows for the existence of other competing monetary assets, bonds and equities, in this system. Keynes poses the question of how the financial system absorbs the reserves and deposits created by the banking system. He argues that rates of return on bonds and equities must adjust until wealth holders are content to hold these assets and deposits in the proportions in which they are being supplied to the public. Thus a change in the reserve policy of the central bank forces a change in the rates of return to bonds and equities.

Since the rates of return on bonds and equity establish the cost of capital funds to firms, changes in these rates of return alter the incentives for firms to make long term investments. A fall in the interest rate engendered by an expansion of bank reserves encourages fixed investment, and this increase in spending by firms raises the level of aggregate demand in the whole economy, normally by a multiple of the initial increase, because households tend to spend part of their additional income as demand expands. In this view there is a close relation between the reserve creation of the central bank and the level of economic activity, mediated by the interest rate on bonds, the price of equities, and the fixed investment policies of firms.

Keynes presents this theory analytically as a correction of Fisher's arguments. First, Keynes insists that the velocity of money, the ratio of desired holdings of money to the value of transactions, responds systematically to the level of interest rates. Second, Keynes argues that prices are not the only variable available to adjust the value of transactions, given the quantity of money and the velocity of money. The other variable is the volume of transactions itself, which changes with the level of economic activity called forth by aggregate demand. While Keynes does not rule out the possibility that price adjustments may, under certain circumstances, be involved in reconciling the value of transactions to the quantity of money and velocity, he deemphasizes this case in arguing that typically the level of economic activity and hence the volume of transactions adapts. Furthermore, Keynes suggests that the relation between money demand, interest rates, and the level of economic activity (in Fisher's terms, the velocity of money) is volatile, subject to sharp changes depending on the mood of wealth holders and their expectations and fears about the future.

Keynes couches his theory in quite traditional terms. He shares with Fisher the concept of a demand for money, or velocity, that establishes a relation between the quantity of money the system will absorb and the levels of prices, interest rates, and economic activity. He also shares with Fisher the procedure of eliminating variables one by one as possible equilibrating factors and arguing that the remaining variable must be the one that adapts. Thus his differences with traditional theory turn on which variable he views as predetermined, and on the emphasis Keynes puts on variations in interest rates as mediating the response of the economic system to changes in the quantity of money. Thus Keynes manages to reverse the classical presumptions that money affects prices but cannot alter the level of interest rates or economic activity, without adopting the view that money is largely endogenous to the economic system.

In historical terms Keynes's theory is a step toward constructing a monetary theory that corresponds to the realities of fully developed industrial capitalism. In his deemphasis of convertibility as a limit on the operations of the central bank Keynes creates a theory that does not depend on the existence of a money commodity. In the place of the traditional emphasis on the money commodity and the relation of domestic money to it, Keynes gives the centre of the stage to the problem of the regulation of aggregate demand and investment. Keynes's vision of the economic system is not that of a self-regulating entity which the economist seeks to understand, but of a complex set of causal linkages that a policymaker seeks to guide.

Keynes's theory of money establishes the framework within which the most influential post-World War II monetary theorists have worked. The basic elements, a demand for money which is a function of income, wealth, and the rates of return on alternative assets, an exogenously given supply of money, and a connection between money and real activity through changes in the rates of return and prices of nonmonetary financial and real assets, serve as the building blocks for both the new quantity of money theory of prices, and extensions of Keynesian theory. But within this framework, different scholars emphasize one or another element to reach quite different policy conclusions.

In the decade after 1945 Keynesian orthodoxy took the position that 'money doesn't matter', in that spending decisions of consumers and firms were largely independent of asset rates of return, and more responsive to expectational variables. This view was supported by the idea that close substitutes for monetary assets could be produced by banks and other financial actors. Thus any attempt to restrict economic activity by limiting the expansion of bank reserves would be circumvented by the substitution of other liabilities. This extreme nonmonetary interpretation of Keynes fell into disfavour as the advanced capitalist countries in the postwar period began to rely more and more heavily on monetary policy as a tool for regulating aggregate demand and the external value of their currencies.

A strong reaction to this deemphasis of monetary factors in the determination of aggregate demand came in Milton Friedman's (1956) resurrection of the quantity of money theory of prices within the Keynesian framework. Friedman argued that as a matter of empirical fact the demand for money is a highly stable function of a small number of relevant variables. He accepted Keynes's idea that the supply of money was exogenously determined by central bank policy, and concluded that changes in the supply of money would have regular and predictable effects on money income and asset rates of return. Friedman also put forward the claim that there are few good substitutes for money (although there has been some uncertainty as to exactly what his theory regards as a monetary asset), so that the demand for money is an inelastic function of the rates of return on competing assets. This implies that changes in the supply of money will be reflected in changes in money income rather than in rates of return. This line of argument leads naturally back to Fisher's conclusion that the level of real economic activity is determined by real factors independent of money, so that the ultimate effect of changes in the supply of money is entirely absorbed by changes in money prices. This series of empirical hypotheses allows Friedman to restore the claims of Fisher's quantity of money theory of prices within Keynes's theoretical framework. Because the new quantity of money theory of prices depends so heavily on empirical claims, it has come under strong questioning as econometricians have attempted to test it with historical data. The demand for money defined in any particular way exhibits more instability than Friedman claimed, and in some definitions a higher elasticity with respect to rates of return on competing assets than is necessary for Friedman's strong conclusions to hold. While it is possible to redefine the monetary aggregate to improve the statistical evidence for the new quantity of money theory of prices, this path opens up potential criticism of ex post theorizing, that is, choosing the definition of the monetary aggregate to save the theory in its confrontation with evidence.

Another pole of Keynesian interpretation is represented by the work of James Tobin (1982).

Tobin also adopts Keynes's conception of a demand for money, but supplements it with demand functions for all other financial and real assets. In this vision money is one part of a spectrum of financial assets, all of which must find their place in wealth holders' portfolios through a mutual adjustment of rates of return and assets prices. For Tobin the possible impacts of monetary changes on economic activity are varied and complex, because they depend on the exact response of the whole structure of rates of return on competing assets to the monetary change, and to the possible reactions of these changes in rates of return on decisions to consume and invest. Tobin accepts Friedman's conclusion that the impact of monetary changes will be absorbed in money prices, but only for a very long run. In the more policy-relevant middle run, there are substantial effects monetary policy can have on the level and direction of economic activity. An expansive monetary policy, by lowering rates of return on bonds and raising the prices of equities, will encourage investment, thus raising the whole level of economic activity, and shifting the emphasis of production toward investment and growth. A contractionary monetary policy, even if it is offset by expansive fiscal policy, so that the overall level of economic activity remains unchanged, will tend to choke off investment and deter long term growth.

These Keynesian lines of thought have been enriched and somewhat modified by incorporating them into models of open economies, in which trade and capital flows are important, as in the work of Robert Mundell (1971). In an open economy with a convertible currency, the supply of domestic money cannot be exogenous. If the central bank expands the supply of money, it will find the public exchanging domestic monetary claims for international reserve assets to offset the expansion. In this context the main scope for monetary policy is at the international level, in the concerted efforts of all the central banks to expand or contract liquidity. In an open economy with a floating exchange rate, and capital markets open to the world, the rates of return on domestic assets will be pegged to world rates of return. In this situation a change in the supply of money has its main effects through changes in the exchange rate. A central bank can influence domestic economic activity in the short run by expanding the supply of money, driving the exchange rate down, and thus expanding the demand for exports. These effects will dissipate over time as domestic price levels adjust, so that the long run conclusions of the quantity of money theory of prices still hold in the open economy framework.

The new quantity theory's claim that in the long run monetary policy cannot affect real economic activity has been transferred even to the short run in the theories of Robert Lucas (1981) that apply the concept of 'rational', or self-fulfilling expectations to simple, stylized macroeconomic models. In this view the public is very quick to learn whatever systematic rule the central bank follows in formulating monetary policy. Once they have learned it, the public will tend to offset the central bank's operations with speculative movements of private portfolios, or through instantaneous price adjustments so as to neutralize any real effects of the policy. Unanticipated or unsystematic changes in the supply of money can affect real economic activity precisely because the public cannot distinguish these changes from changes in the underlying parameters of tastes, technology and resources that are thought to determine real decisions. Thus money itself can have short run effects on economic activity, but the rational expectations school argues that these possible effects can never be used for policy ends in a systematic fashion. It is unclear how general these results are, especially in circumstances where there are important differences in information and beliefs in different segments of the public, and where costs of learning the true structure of the economy (if such a structure actually exists) are significant.

The research of Don Patinkin (1965) and Kenneth Arrow and Frank Hahn (1971) on the insertion of money into fully specified general equilibrium theory has yielded some interesting clarifications of old arguments, but has not been able to reach sharp and sweeping conclusions like those of the new quantity of money theory of prices. By treating real balances of monetary assets as another good symmetrical with produced

and consumed goods, Patinkin has shown that out of equilibrium the stock of real balances in principle affects the demands and supplies of all other assets. This argument is fatal to Fisher's simple procedure of separating the determination of relative prices and of the level of money prices. Hahn points out the paradox involved in assuming that money (as a thing, now, not a social relation) is valued only for having a positive price. In general in any monetary general equilibrium economy there exist equilibria in which money has a zero price, that is, a nonmonetary equilibrium. Since the non-monetary equilibrium is quite different from the monetary equilibria, and may involve much lower levels of trade and production, this abstract observation leads to a qualitative understanding of the role of money in facilitating economic activity. This general equilibrium modelling generally accepts the framework of the new quantity of money theory of prices in positing the existence of a single, given, monetary asset with no close substitutes, and in abstracting from the questions of how monetary liabilities come to exist, and whether or not they can be produced by private agents.

Hyman Minsky (1982) puts forward, in contrast, a theory of the relation of money to economic activity in which qualitative changes in the private issuance of monetary liabilities plays a central role. In Minsky's view, firms issue liabilities to finance production based on uncertain (and not necessarily self-fulfilling) expectations about future profitability. As an economic expansion develops, these expectations become more buoyant, and more liabilities are issued. This process gradually erodes the quality of the liabilities, because there comes to be a larger and larger probability that profit realizations will not in fact allow all the commitments to be met. Each firm tends to move towards thinner and thinner margins of equity in its financial position; firms that are reluctant to follow this policy find themselves severely punished competitively in the short run. The deterioration of the quality of liabilities sets the stage for a financial crisis, in which many firms face difficulties in meeting their commitments, and new lending is extended only on much tougher terms. In the absence of State intervention to substitute its liabilities in part for lower quality private liabilities, the resulting collapse of the financial system has strong repercussions on levels of economic activity as firms find it difficult to finance new productive outlays. Minsky's account emphasizes the qualitative, rather than the purely quantitative effects of monetary liabilities on economic activity. It also goes beyond quantity of money theories in seeing the space of monetary liabilities as constantly shifting in its properties, as new liabilities are invented and old ones take on a different function with the development of production. In the place of a single, inelastically supplied, monetary liability with known and unchanging properties, the spectrum of financial assets in Minsky's view is filled up with elastically supplied liabilities of unknown and constantly changing properties.

Channels of influence run both from money to economic activity and from economic activity to money. Whether money takes the form of a commodity produced by the system, or of liabilities issued to finance production, the creation of monetary assets is an incident in the cycle of production. But it is at least partly through the availability and cost of finance that levels of planned production are determined, and confined within the productive capacities of the whole society as Michal Kalecki (1971) has emphasized. Different monetary theories have emphasized one or another side of this mutual interaction, without reaching a fully adequate synthesis.

The relation between money and economic activity must be analyzed in explicitly dynamic terms because monetary and financial institutions constitute an important feedback loop in commodity-producing economies. The properties of the equilibria of a system often fail to reveal its dynamic behaviour. In equilibrium situations the powerful forces running from money to economic activity are balanced by those running the other way, and monetary effects tend to disappear from view. The contemplation of such equilibrium situations is an insufficient guide to understanding the effects of money on economic activity in general.

See Also

- ▶ Circular Flow
- ▶ General Equilibrium
- ▶ Money and General Equilibrium Theory
- ▶ Neutrality of Money
- ▶ Quantity Theory of Money

Bibliography

Arrow, K.J., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.
Fisher, I. 1911. *The purchasing power of money*. New York: Macmillan.
Friedman, M. (ed.). 1956. *Studies in the quantity theory of money*. Chicago: Chicago University Press.
Hume, D. 1752. In *Writings on economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.
Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy*. Cambridge: Cambridge University Press.
Keynes, J.M. 1936. *The general theory of employment, interest, and money*. London: Macmillan.
Lucas, R.E. 1981. *Studies in business cycle theory*. Cambridge, MA: MIT Press.
Marx, K. 1867. In *Capital*, vol. I, ed. F. Engels. New York: International.
Minsky, H. 1982. *Can 'It' happen again?* Armonk: Sharpe.
Mundell, R. 1971. *Monetary theory*. Pacific Palisades: Goodyear.
Patinkin, D. 1965. *Money, interest and prices*, 2nd ed. New York: Harper & Row.
Ricardo, D. 1811. *The works and correspondence of David Ricardo*. Vol. III, *Pamphlets and papers 1809–1811*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan. London: Methuen, 1961.
Tobin, J. 1982. *Essays on economics: Theory and policy*. Cambridge, MA: MIT Press.

Money Supply

Benjamin M. Friedman

Abstract

Governments supply money not only for use in everyday transactions but also, in the modern era, in order to influence their economies. In most advanced industrialized economies the demand for money is sufficiently unstable to make the quantity of money supplied, or its growth rate, an unreliable guide to how monetary policy influences either prices or real economic activity. Most central banks therefore set a designated interest rate, not the quantity or growth of money supplied. But because money supply and money demand help determine market interest rates, the money supply process remains essential to analysing how monetary policy operates.

Keywords

Aggregate demand; Central bank independence; Central bank reserves; Central banking; Central banks; Friedman, M.; Goodhart's Law; Hyperinflation; Inflation; Interest rates; Monetary base; Monetary policy; Monetary policy rules; Money; Money demand; Money multiplier; Money supply; Open market operations; Optimal monetary policy; Reserve-to-deposit ratio

JEL Classifications

E5

Supplying money for use in everyday transactions, so as to obviate the need for cumbersome barter, has been a function of governments for more than 2,000 years. Not surprisingly, government-issued money, once in existence, rapidly became a store of value as well. As an aspect of the history of human society and institutions, the process by which governments supply money has naturally attracted substantial attention. But the primary interest in money supply within the discipline of economics has stemmed from the proposition that movements in money are an important –according to some views, the most important – determinant of movements in prices, in output and employment, and in other economic phenomena of well-established interest on their own account.

Two analytical frameworks that rose to prominence in the latter half of the 20th

century – indeed, that dominated macroeconomic thinking during much of that period – attached just this importance to money: quantity-theory monetarism, and IS–LM Keynesianism. Both these frameworks, however, took for granted that governments conduct their affairs (specifically in this context, that central banks conduct monetary policy) in such a way as to create independent movements in the supply of money, as opposed to merely passive movements in response to changes in money demand that therefore could not plausibly be the cause of movements in either prices or real economic activity. As of the outset of the 21st century, however, the number of central banks that in fact carry out their responsibilities in such a way is small and shrinking. Instead, most central banks implement monetary policy by setting some designated short-term interest rate.

As a result, interest in how money is supplied has sharply diminished among economists, and the details of the money supply process are now often omitted from the standard economics curriculum. (Examples at the graduate level are the instructional text by David Romer 2006, and the theoretical treatise by Michael Woodford 2003). In the absence of some substantive knowledge of how money is supplied, however, just how a central bank can set 'the interest rate' would remain mysterious. Even if the number of central banks that actively seek to influence money supply as an element of the conduct of monetary policy shrinks to zero, therefore, money supply is unlikely to disappear from the purview of economics altogether.

The Analytical Basics

The first recognized monies supplied by governments for ordinary economic use mostly consisted of precious metals. The authorities' role was to provide standardized units, together with what amounted to stamped certification that the amount of metal in the coin or other object conformed. Apart from the certification, therefore, anyone who had an adequate quantity of the chosen metal could supply money along with the government.

In the more modern conception of money supply, relevant only since the 19th century, money is a form of debt. Most government-issued money consists of currency, which represents the liability of a partly or wholly government-owned central bank. Currency is typically not interest-bearing, and so the motives for holding it do not stem from its role as an earning asset. And although it is the government's (the central bank's) liability, in modern times it usually does not represent an obligation on the government's part to pay the bearer in some other form. Instead, both private citizens and businesses hold these government liabilities for their convenient use in everyday transactions, normally enforced by their statutory status as legal tender.

The fact that government-issued money is supplied as the liability of the central bank, and the presumption that the central bank has control over its balance sheet, together create the conceptual foundation for viewing the supply of money as a tool of economic policy. Indeed, much of the initial interest in this subject in the modern era arose from the experience of countries where the central bank had lost control of its balance sheet for some period of time, often in the aftermath of war or under other circumstances that prevented the government from raising ordinary revenues to cover its ongoing expenditures. The observation that such episodes often led to spiralling hyperinflation, with rising prices requiring the government to issue more money (in the absence of other revenues) and the larger supply of money leading to further increases in prices, immediately suggested a connection between money supply and prices, if not real economic activity as well.

Apart from situations of runaway money supply and hyperinflation, however, the issuance of currency is usually not the focus of economists' interest in how the supply of money relates to economic activity. While the great majority of government-issued money in the economically advanced countries now consists of currency held by the public (as of 2006, 69 per cent in the United Kingdom, and 95 per cent in the United States), currency is nonetheless only a small part of the money that individuals and firms use for savings and to execute everyday economic

transactions. The money that individuals and firms use mostly consists of deposits issued by banks and other financial institutions. In the United Kingdom, deposit money outweighs currency by more than 30 to 1. Even in the United States, where the country's currency is also commonly used in both legal and illegal transactions around the world, the ratio is more than 8 to 1. Moreover, although in principle a central bank could seek to influence the economy by manipulating how much currency it supplies, in practice most central banks supply currency passively to accommodate whatever demands the public may have. (The role of currency issuance as a source of government finance – the heart of most examples of hyperinflation – is likewise limited in most economically advanced countries. Even in the United States, with demand for the currency enlarged by the use of US dollars in other countries, issuance of currency in a typical year amounts to only one to two per cent of the federal government's spending.) The simple construct of an economy in which the public depends entirely on government-issued currency to execute economic transactions, and the central bank exerts its economic influence by expanding or contracting the supply of that currency, is a textbook instructional device with limited relevance to most actual economies.

From the perspective of any active connection to either nonfinancial economic activity or the pricing of assets in the financial markets, therefore, what matters is the larger money supply issued by banks and other depository institutions (hereafter simply 'banks' for short). And in most modern banking systems, what gives the central bank the ability to influence the volume of deposits that banks in the aggregate create is its control over the amount of its own liabilities that it supplies for banks to hold. While most of the central bank's liabilities consist of currency held by the public, the remainder (31 per cent in the United Kingdom, and only five per cent in the United States, as of 2006) are held as assets – normally called 'reserves' – by the banks. The link between the banks' creation of deposits for the public to hold and their own holdings of reserves at the central bank constitutes the heart of the money supply process for purposes of a connection to most matters of concern to monetary policy.

Banks hold central bank reserves – and, importantly, hold more reserves as they have more deposits outstanding (all other things equal) – for several reasons. First, in traditional 'fractional reserve' banking systems, banks are required by law to hold such reserves in amounts equal to at least some fixed percentage of their outstanding deposits. Hence a larger supply of reserves makes it possible for the banks to do more lending (or buy more securities) and therefore create more money. Conversely, contracting the supply of reserves requires banks to shrink the amount of deposits they have outstanding, normally by not extending new loans to replace existing credits that mature or are otherwise repaid, or by selling securities.

Second, banks need a supply of currency to satisfy customers who draw on their accounts or present checks or other negotiable instruments for payment. In some banking systems, currency held by banks (as opposed to currency held by the public) is counted as part of banks' reserves. When a customer cashes a check, therefore, bank reserves fall and there is a corresponding increase in currency held by the public. (Because the central bank is not a party to the transaction, the total amount of central bank liabilities remains unchanged.) But banks cannot satisfy such demands unless they are holding an adequate amount of currency to begin with. And the greater the bank's volume of business, including in particular the amount of deposits it has outstanding against which its customers may want to draw, the more currency – hence the more reserves, if bank-held currency counts as reserves – the bank will ordinarily hold.

Third, banks also need to settle transactions with one another. If a customer of one bank deposits a check written against an account at another bank, the two banks must transfer some asset from one to the other. The same is true if one bank sells a security to another. Although banks in most countries have various mechanisms, like private clearing houses, for effecting such transfers without involving the central bank, some

inter-bank transactions do normally settle by transferring reserves at the central bank from the paying bank to the receiving bank. In order to participate in that process, banks therefore need to hold at least some amount of reserves; and the more deposits the bank has outstanding, the more inter-bank transactions it may have to settle on a given day, and so the more reserves it will ordinarily hold. Moreover, in some banking systems the central bank reinforces the demand for its reserves by requiring banks to settle certain classes of inter-bank transactions in this way. Especially in systems where there are no reserve requirements in the traditional form of a stated minimum percentage of outstanding deposits, requiring the banks to settle inter-bank transactions in this way reinforces the banks' need to hold central bank reserves.

Banks' demand for reserves, therefore, is in many ways analogous to the public's demand for money. Reserves provide banks with an ability to do business, just as the money that individuals and nonfinancial firms hold enables them carry out their everyday economic affairs. That ability has value, but not infinite value. Hence the more expensive it is for banks to hold reserves, in terms of interest forgone by holding reserves instead of some other asset, the more banks will seek to economize on their reserve holdings in relation to their outstanding volume of deposits. For a given amount of deposits, therefore, banks' demand for reserves is negatively elastic with respect to the interest rate on alternative assets (typically loans or securities), just as the public's demand for money is negatively interest elastic for a given amount of income being earned or transacting being done. If reserves at the central bank bear an interest rate that varies in close step with what banks can get from holding other earning assets, this negative interest elasticity is likely to be small, or even trivial. But if the interest rate that the central bank pays on reserves is fixed (in the United States, for example, it is fixed at zero), or even if it varies together with market returns but only imperfectly, the negative interest elasticity in banks' reserve demand is likely to be significant. (The classic paper making this point is Dewald 1963.)

The analytical mirror image of banks' negatively elastic demand for reserves, for a given volume of deposits outstanding, is their positively elastic willingness to create deposits for a given amount of reserves that they hold. The higher are market interest rates on earning assets, compared to whatever rate the central bank pays on reserves, the greater is the incentive for banks to stretch their reserves further by making more loans and buying more securities – and in the process creating more deposits – rather than leaving an increasingly expensive cushion of reserves that may provide benefits (less risk of having to take abrupt action in the event of a shortfall, for example) but are costly nonetheless.

The result is a positively interest-elastic supply of money, representing the behaviour of banks, to go along with the usual negatively interest-elastic demand for money representing the behaviour of the households and firms that hold bank deposits, together with currency, as the money that they use for economic purposes. In the absence of some pathology, the intersection of this positively interest-elastic money supply and negatively interest-elastic money demand determines the equilibrium quantity of money created and held, for a given supply of reserves and a given level of income, together with the interest rate at which the market clears.

(And, because the positively interest-elastic supply of money is simply the mirror image of the negatively interest-elastic demand for reserves – both represent the same aspect of banks' behaviour – the market for reserves is likewise in equilibrium, with demand equal to whatever quantity of reserves the central bank is supplying, at the same interest rate.) Integrating this partial equilibrium of the money market (and the reserves market) with the demand for goods and services then completes a simple representation of the economy's aggregate demand. Further integrating that aggregate demand representation with aggregate supply, importantly including the labour market, in turn completes the economy's short-run general equilibrium (short-run in that such dynamic elements as the stocks of capital, technology, and other relevant factors are still unaccounted for).

In some treatments of money supply within the economics literature, this explicit supply–demand equilibrium in the markets for money and reserves is, instead, implicitly represented by a simple 'money multiplier' stating the relationship between the total liabilities supplied by the central bank – often called the 'monetary base' – and the resulting amount of money, including bank deposits as well as currency. Purely as a matter of arithmetic, specifying the ratio of reserves to deposits that the banks choose to hold (influenced in part by whatever reserve requirements and other institutional strictures banks face), and the ratio of currency to deposits that the public chooses within its holdings of money, is sufficient to determine the quantity of money that goes along with any given monetary base set by the central bank. But the banks' reserve-to-deposit ratio depends in part on interest rates as well, and the public's demand for currency often varies with a host of factors (confidence in the banking system, use of currency abroad or for purposes of illegal transactions, and so on), so that the 'money multiplier' representation is really just a shorthand simplification that works well or badly depending on the strength of the relevant interest elasticities and the extent of variation in interest rates and the many other factors involved. (See, for example, Cagan 1965. A brief statement of the central ideas appeared in Friedman and Schwartz 1963, ch. 2, sec. 4). Underneath, the supply–demand equilibrium established by the central bank's supply of reserves, banks' behaviour in demanding reserves and supplying deposits, and the public's behaviour in demanding both deposits and currency, is what establishes an economy's money supply. (For a fully articulated treatment, see Modigliani et al. 1970.)

The Link to Monetary Policy

The logical starting point in this process is the central bank's supply of its own liabilities, and it is the central bank's control over the liabilities it issues that gives the supply of money its place in economic policy. Until fairly recently – well into the 19th century – governments issued either coins or paper currency mostly as a means of payment for goods and services they purchased. Such actions were, in effect, a combination of what have come to be known as fiscal and monetary policies. In the modern era, however, especially with the advent of central banks as distinct and often quasi-independent governmental institutions, economists have thought of fiscal and monetary policies as likewise distinct.

In the absence of a securities market, or some similar set of financial institutions, it is difficult to conceive of how monetary policy would operate independently of fiscal policy: how could the government, in such a setting, increase the amount of money outstanding without simultaneously making either a purchase or at least a transfer payment? One metaphor sometimes used in the theoretical economics literature to represent such an action – and which only serves to indicate how far-fetched such a situation is – is to picture the government dropping money from a helicopter. While monetary and fiscal policies are distinguishable in most modern economies, central banks, of course, do not drop money from helicopters. The reason is that the economies in which they operate in fact have securities markets.

The primary means by which central banks in most modern economies change the amount of their liabilities outstanding is to purchase, or sell, securities – actions typically called 'open market operations'. When the central bank buys a security, it makes payment by increasing the amount of reserves credited to the seller's bank. (In systems in which bank-held currency is counted as part of reserves, the consequence is the same even if the central bank makes payment by delivering currency to the seller's bank.) When the central bank sells a security, it correspondingly receives payment by reducing the amount of reserves credited to the buyer's bank. In either case, the central bank's assets, consisting mostly of the securities it holds, and its liabilities, consisting partly of the reserves credited to banks, rise or fall in lockstep. But because of the ways in which banks' ability to create deposits depends on their holdings of reserves, the change is not economically irrelevant. Changes in the supply of reserves, effected via open market operations, shift a key

underpinning of the equilibrium in the reserves market and the money market, thereby changing not only the resulting quantity of money but the yields and prices of non-money assets and ultimately the equilibrium of the nonfinancial economy as well.

Not all open market operations carried out by central banks change the quantity of reserves. Most importantly, the central bank also needs to accommodate the public's changing demand for currency. In a growing economy with rising prices, the demand for currency is usually increasing. When individuals and businesses go to their banks to get more currency, their doing so increases the amount of currency in public circulation but reduces the amount of the banks' reserves (as long as bank- held currency is counted as reserves). As a part of their normal ongoing procedures, therefore, most central banks routinely purchase securities – that is, carry out open market operations – in order to offset such reductions in reserves due to increasing public demand for currency. Central banks also regularly carry out open market purchases or sales in order to prevent short-run fluctuation in other technical factors, such as international transactions and variations in the amount of checks currently in the clearing process, from affecting the supply of reserves.

Central banks can also create reserves by lending to banks, rather than buying earning assets from them, and in some countries' systems the lending of reserves is more important for purposes of carrying out monetary policy than open market operations. Whether banks distinguish between reserves that they have borrowed from the central bank and reserves that they simply own outright (often called 'nonborrowed reserves' to distinguish the two) depends on the specifics of the individual system's institutions. Most obviously, borrowed reserves are a liability of the bank, on which it presumably has to pay interest, while its nonborrowed reserves are an asset on which it may or may not earn interest. In addition, in some systems (the United States, for example), borrowing reserves from the central bank exposes a bank to regulatory oversight with implicit costs well beyond what the interest rate paid would suggest.

Whether reserves are borrowed or nonborrowed, however, the essence of monetary policy is the central bank's provision of reserves to the banking system.

The recognition of the way in which that role played by the central bank potentially affects an economy's money supply, interest rates, asset prices, nonfinancial activity, and prices and wages, in turn sets the stage for both normative and positive consideration of monetary policy. The ensuing economics literature has become vast. In most countries the corresponding public discussion is likewise active and intense.

The modern economist most identified with emphasizing the role of money supply in the conduct of monetary policy – as opposed to focusing on interest rates, or measures of reserves in the banking system, or other relevant indicators of what a central bank is doing in this respect – is Milton Friedman. At the most fundamental normative level, Friedman advocated a long-run policy of shrinking the supply of money (by which he meant government-issued money) at a rate adequate to render nominal interest rates on assets closely substitutable for money equal to zero on average over time. The basic logic was that, since the government could create such money at essentially no cost, it should be costless for the public to hold; the public's effort to economize on holdings of money balances, when market interest rates on money substitutes are positive, represents a deadweight loss to the economy (see Friedman 1969). Given the demonstrated dangers of deflation, however – with a positive real rate of interest, negative inflation would be necessary to achieve a zero average nominal interest rate – this recommendation had little impact on actual monetary policy.

At a more practical level, however, over short- and medium-run horizons Friedman advocated keeping the supply of money (by which he meant the deposits and currency held by the public) growing at a constant rate. Here the argument was that the influence of monetary policy on both prices and real economic activity operates with lengthy delays, subject to unpredictable variation, and that active attempts by the central bank to use monetary policy to offset nonmonetary influences

on the economy were likely to be destabilizing (see Friedman 1953, 1956). Many other economists, more optimistic about the prospects for using active variation in monetary policy to blunt the influence on the economy of factors that the central bank could either foresee or at least recognize quickly once they had occurred, followed Friedman in advocating the use of growth in the money supply as the way to gauge whether the central bank was exerting a stimulative or a contractionary force on economic activity. Beginning in the 1960s, but more so in the 1970s, many central banks around the world implemented these recommendations by adopting one or another form of explicit target for the growth of its money supply.

The Role of Empirical Evidence

The crucial empirical underpinning of such policy frameworks, whether they involved constant money growth or attempts at active stabilization nonetheless benchmarked by money growth, was the observation that movements in money bore a reliable relationship to movements in income and prices. Early in the post-Second World War period, Philip Cagan documented such a relationship between money growth and price inflation in several well-known episodes of hyperinflation in Europe that had followed each of the two world wars (Cagan 1956). But hyperinflation in the context of post-war chaos (especially for the war's losers) bore only limited implications for the conduct of monetary policy under more normal circumstances. In a massive historical study, Milton Friedman and Anna J. Schwartz documented the relationships between money and prices, and also money and income, for the United States during the period 1867–1960 – including the Great Depression of the 1930s but also many more ordinary business fluctuations as well – and following their work many other empirical researchers attempted similar (though mostly smaller-scale) studies for other countries and other time periods (Friedman and Schwartz 1963).

At the conceptual level, the central idea linking this empirical research to the implied role of money supply in conducting monetary policy was that, if fluctuations in money growth and fluctuations in income and/or prices are systematically related, and if the observed fluctuations in money growth within those relationships represent independent movements of money supply, then the central bank can exploit those relationships by purposefully steering the money supply along an optimally chosen course (which may or may not be a simply constant-growth path). Following the work of Friedman and Schwartz, and the many other researchers who applied ever more sophisticated empirical methodologies to the same line of enquiry, questions about each of these two underlying issues – how strong the observed relationships are, and whether they result from independent movements of money supply – generated a similarly large literature.

One immediate difficulty, recognized early on, is that, since money supply necessarily equals money demand, inferences about the money–income or money–price relationship on the basis of observed movements in money are subject to the usual problem of statistical identification. (An early paper making this point was Teigen 1964. Another, addressed more explicitly to the work of Friedman and Schwartz, was Tobin 1970.) Hence what may look like a relationship between movements of prices and income induced by movements in money supply may in reality be movements in money demand induced by movements in prices and income. Further, unless the central bank takes its decisions affecting money supply with no regard for the behaviour of prices and income, the observed relationships may also represent the reactive behaviour of the central bank itself. Indeed, under some plausible accounts of how central banks make monetary policy, relationships of the kind observed in the data would spuriously emerge. (An early paper making this point was Goldfeld and Blinder 1972.) Still more fundamentally, even if the relationships observed between money and either income or prices actually did represent exactly the kind of causal influence of money supply that was claimed, the attempt by the central bank to exploit such a relationship for policy purposes, once widely recognized, could

cause the relationship to change or even break down altogether. (The classic statement of this proposition in a general context is Lucas 1976. For a formulation in the specific context of monetary policy, see Goodhart 1984; the original formulation of 'Goodhart's Law' dates to 1975 when this paper was first presented.)

Starting in the mid-1970s, however, and then increasingly so over the next two decades, these questions became moot. Fluctuations in money growth no longer appeared to bear much observed relation to fluctuations in either income or prices over time horizons that were useful for conducting monetary policy, especially after controlling for other obvious information like past movements of income and prices themselves. In parallel, the evidence indicated that money demand was unstable. The presumption of a stable functional relationship between money demand and income or prices had always been central to the claim that money supply was a useful tool for purposes of monetary policy. But now evidence for a stable money demand gave way, in one country after another, to evidence of instability.

The reasons for the disappearance of stable money demand were many, and, at a qualitative level, straightforward to understand. (The empirical money demand literature is a separate subject; for a survey, see Goldfeld and Sichel 1990. For an earlier survey, written before the instability became so widespread or so evident, see Laidler 1977, ch. 7.) One reason was changing regulation (in the United States, for example, the removal of the prohibition against banks' paying interest on checkable deposits, and also of the ceilings limiting the interest that banks could pay on interest-bearing savings deposits). Another, in part prompted by regulatory changes, was innovation in the kinds of deposits and deposit-like instruments that banks and other financial institutions offered their customers (for example, money market mutual funds). A third was the electronic revolution, which made various forms of financial transactions ever easier and less costly (for example, shifting funds between checkable and non-checkable accounts). A fourth was rapid globalization, which made businesses in particular, but many individuals as well, increasingly willing to hold assets, and to borrow, in multiple currencies, and to substitute readily among them. But regardless of the precise reasons, which presumably varied from one country to another, money demand no longer appeared to be stable. Nor, in parallel, did the relationships of a simpler form between money and either income or prices that had spurred policy interest in money supply in the first place.

The Decline of Money Supply as a Tool of Monetary Policy

In the absence of empirical evidence of stable money demand, the rationale for the role of money supply as a tool of monetary policy collapsed as well. If money demand is unstable, then even perfectly stable money supply introduces into income and prices the influence of whatever disturbances to the public's money-holding behaviour occur. Under those circumstances, the central bank can do a better job of stabilizing either prices or income, over the short or medium run, by fixing some interest rate and thereby allowing fluctuations in money supply to accommodate fluctuations in money demand that occur for reasons unrelated to movements of income and prices. (The classic paper making this point is Poole 1970; for a survey of the optimal monetary policy literature along these lines, including the role of money supply behaviour along with money demand, see Friedman 1990. In the long run, however, there must be at least some absolute nominal element in the policy mechanism to anchor the price level; the interest rate is a relative price, not an absolute price.) Following the increasing evidence of money demand instability, and the collapse of money–price and money–income relationships, that is precisely what an increasing number of central banks have done.

The experience in the Unites States is illustrative. The Federal Reserve System, the US central bank, first began to take explicit note of money supply movements in formulating its monetary policy in 1970. In 1975 the US Congress adopted a resolution requiring the Federal Reserve to

announce, in advance, quantitative targets for the growth of key money (and credit) aggregates and, after the fact, to report to the relevant Congressional oversight committees on its success or failure in meeting these targets. In 1979 the Federal Reserve publicly declared an intensified dedication to controlling money growth, with the main focus on the narrow M1 aggregate (consisting primarily of currency and checkable deposits), and adopted new day-today operating procedures, centred on the supply of nonborrowed reserves, designed to enhance its ability to achieve control of M1.

The movement towards ever greater emphasis on money supply in US monetary policy took less than a decade; unwinding it took only a little longer. In 1982, the Federal Reserve recognized the increasing instability of demand for M1 and shifted its focus to the broader M2 (including not only currency and demand deposits but also most forms of time and savings deposits). Soon thereafter, it abandoned its operating system based on nonborrowed reserves, in favour of simply setting the federal funds rate (the overnight interest rate on bank reserves) at the level most likely to achieve the desired M2 growth. After 1986 the Federal Reserve stopped setting a target for M1 growth, but continued to do so for M2 and M3 (a still broader aggregate). In the late 1980s evidence based on how the Federal Reserve changed the federal funds rate in response to observed movements of money suggested that the M2 growth target still bore significant influence on US monetary policy. (See, for example, Friedman 1997; but the empirical literature on this issue is voluminous.)

That influence had mostly dissipated by 1990, and in 1993 the Federal Reserve publicly 'downgraded' the role of its M2 target. Thereafter it continued to set 'ranges' for M2 and M3 growth, but it made clear that these were not actual money growth targets; they were merely 'intended to communicate its expectation as to the growth of these monetary aggregates that would result' under specified assumed conditions. In 1998 the Federal Reserve further confirmed that these ranges were not 'guides to policy'. In 2001 it stopped setting such ranges altogether.

The pattern in most other countries was roughly parallel. By 1980 the use of money supply targets for monetary policy was an idea whose time had come. Most of the major central banks had put such targets at the core of their policymaking process. By 1990 money growth targets were already largely a thing of the past. By the mid-1990s most central banks had either de-emphasized such targets or dropped them altogether. By 2000 it had become standard that central banks carry out monetary policy by setting some short-term interest rate. Money supply mostly disappeared from public discussion, and the professional economics literature largely dispensed with the now-unnecessary apparatus of money demand, money supply, and likewise demand and supply in the market for reserves. (See, for example, Clarida et al. 1999.)

Implicitly, however, that conceptual apparatus nonetheless stands behind the ability of central banks to set the designated interest rate in the first place. In principle, a central bank – or anyone else with large enough resources, for that matter – could fix the price or yield on any asset simply by buying or selling that asset in sufficient volume to shift the entire market equilibrium, ultimately including the real returns established by the fundamental economic forces of thrift and productivity. (Given the lags with which monetary policy influences price inflation, in the short run the interest rate the central bank is setting is a real interest rate.) But in fact most central banks normally move the interest rate they use for monetary policy purposes by executing only very small transactions, and in an increasing number of cases they do so without executing any transactions at all; often the mere announcement of what the central bank would like the designated rate to be is sufficient.

What gives a central bank the ability to do so is, presumably, market participants' knowledge that the interest rate being set is closely tied to that on the central bank's own liabilities (in systems like that in the United States, it is exactly that rate), and that the central bank can make the supply of those liabilities whatever it chooses. But market equilibrium requires that the demand for those liabilities equal the supply, and the demand for central

bank liabilities in turn is an aspect of the same behavioural process that determines the supply of money. Hence money supply remains a part of the story, even if now mostly a hidden one.

See Also

- ▶ Friedman, Milton (1912–2006)
- ▶ Inside and Outside Money
- ▶ Monetary and Fiscal Policy Overview
- ▶ Monetary Policy, History of
- ▶ Monetary Transmission Mechanism
- ▶ Money

Bibliography

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

Cagan, P. 1965. *Determinants and effects of changes in the stock of money, 1875–1960*. New York: NBER.

Clarida, R., J. Gali, and M. Gertler. 1999. The science of monetary policy: A new Keynesian perspective. *Journal of Economic Literature* 37: 1661–1707.

Dewald, W.G. 1963. Free reserves, total reserves and monetary control. *Journal of Political Economy* 71: 141–153.

Friedman, M. 1953. The effects of a full-employment policy on economic stability: A formal analysis. In *Essays in positive economics*. Chicago: University of Chicago Press.

Friedman, M. 1956. The quantity theory of money: A restatement. In *Studies in the quantity theory of money*. Chicago: University of Chicago Press.

Friedman, M. 1969. The optimum quantity of money. In *The optimum quantity of money and other essays*. Chicago: Aldine.

Friedman, B.M. 1990. Targets and instruments of monetary policy. In *Handbook of monetary economics*, ed. B.M. Friedman and F. Hahn, vol. 2. Amsterdam: North-Holland.

Friedman, B.M. 1997. The rise and fall of money growth targets as guidelines for U.S. monetary policy. In *Towards more effective monetary policy*, ed. I. Kuroda. London: Macmillan.

Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Goldfeld, S.M., and A.S. Blinder. 1972. Some implications of endogenous stabilization policy. *Brookings Papers on Economic Activity* 1972 (3): 585–644.

Goldfeld, S.M., and D.E. Sichel. 1990. The demand for money. In *Handbook of monetary economics*, ed. B.M. Friedman and F. Hahn, vol. 2. Amsterdam: North-Holland.

Goodhart, C. 1984. Problems of monetary management: The U.K. experience. In *Monetary theory and practice: The U.K. experience*. London: Macmillan.

Laidler, D.E. 1977. *The demand for money: theories and evidence*. 2nd ed. New York: Harper & Row.

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

Modigliani, F., R. Rasche, and J.P. Cooper. 1970. Central bank policy, the money supply and the short-term rate of interest. *Journal of Money, Credit and Banking* 2: 166–218.

Poole, W. 1970. Optimal choice of monetary policy instruments in a simple stochastic macro model. *Quarterly Journal of Economics* 84: 197–216.

Romer, D. 2006. *Advanced macroeconomics*. 3rd ed. Boston: McGraw-Hill/Irwin.

Teigen, R.L. 1964. Demand and supply functions for money in the Unites States: Some structural estimates. *Econometrica* 32: 467–509.

Tobin, J. 1970. Money and income: Post hoc ergo propter hoc? *Quarterly Journal of Economics* 84: 301–317.

Woodford, M. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.

Money Supply in the American Colonies

Farley Grubb

Abstract

The British colonies in North America experimented with legislature-issued paper monies to supplement their specie monies which were in chronic short supply. These experiments were designed to produce inside monies that, unlike specie monies, could not profitably be exported. The nature of British regulation, while leaving room for some variation, constrained the colonies to issuing fiat currencies that were typically tied to paying the future taxes and other dues levied by the issuing colonies. After some early failures, most of these experiments performed well over the quarter century before the Revolution as

revealed by the presence of long-run price stability.

Keywords

Backing or asset theory of money; Barter; Commodity money; Fiat currency; Benjamin Franklin; Inside money; John Law; Purchasing power parity; Quantity theory of money; Real bills doctrine

JEL Classification

N11

The money supply in the British North American colonies was a complex mixture of colonial legislature-issued inside fiat paper monies and outside specie monies. Gold and silver coins (specie) were the principal local and international monies of exchange for Europeans. In British North America such monies could only be acquired through trade or government transfers as gold and silver were not yet mined there. Specie was acquired through trade surpluses with Spanish America and the Caribbean. In addition, British military spending, especially during King George's War (1744–1748) and the Seven Years' War (1756–1763), injected specie into colonial economies. This specie, however, quickly flowed out to cover the colonies' trade deficits with Britain. The British government used mercantilist policies to prevent specie outflows from Britain and encourage specie inflows by holding their colonies in a state of chronic trade deficit with the mother country. As specie passed through the colonies it could only serve in a limited capacity as a local medium of exchange. Given the frequent disruptions to trade flows caused by wars, weather shocks and, in the decade before the Revolution, political boycotts, periods of specie dearth and glut in the colonies were not uncommon. Colonists often complained of a lack of specie. As a result, extensive barter systems using merchant book credit and non-specie commodity monies, such as tobacco, developed to support local exchange. These barter systems were never completely displaced by monetized exchange during the colonial era (Brock 1975; Grubb 2004, 2008; McCusker 1978; Mossman 1992; Rabushka 2008).

Ironically, the relative efficiency of colonial barter was likely responsible for the chronic scarcity of specie. Individual colonists could not capture the positive externalities resulting from the lowered transaction costs in all subsequent local trades that used their specie (as nineteenth-century banks could by loaning banknotes fractionally backed by specie reserves), and they gained more by quickly exporting their specie to buy British goods. Exchanging specie for an entire British good was more welfare-enhancing than the gain from lowered transaction costs in using that specie to acquire a good in local exchange, relative to using barter to obtain that good. As such, the colonies could only gain the lowered transaction costs that a monetary medium of exchange could bring if they could create a medium that could not to any great advantage be exported – a pure inside money. Colonial experiments with paper money, which were also called bills of credit, can be understood in this light.

During the eighteenth century the British North American colonies became the first western economies to rely on legislature-issued fiat paper monies as their principal internal media of exchange. These monetary experiments were neither uniform nor coordinated across the colonies. They were instituted piecemeal – at different times with different motivations and goals. Their institutional structures and relative performances varied as well. These experiments, while wide-ranging, were constrained by British regulations that effectively limited the colonies' rights and abilities to mint their own coins, institute capital controls to limit specie exports, and form private corporations that could effectively function as banks using specie as reserves to support private paper money emissions (which dominated nineteenth-century US paper money creation) (Brock 1975; Grubb 2006, 2008; Mossman 1992; Rabushka 2008).

The Massachusetts legislature in 1690 was the first to issue paper money, followed by South Carolina in 1703, New York, New Jersey, and New Hampshire in 1709, Rhode Island and Connecticut in 1710, North Carolina in 1712,

Pennsylvania and Delaware in 1723, Maryland in 1733, and finally Virginia and Georgia in 1755. In the first eight cases, paper money was created as a solution to the short-run fiscal crises caused by emergency military spending during King William's (1689–1697) and Queen Anne's (1702–1713) wars. Emergency spending during the Seven Years' War led Georgia and Virginia to create their paper monies. For paper money to become a permanent medium of exchange in the peacetime economies of these colonies was not necessarily the motive behind these experiments, although for many colonies it evolved in that direction (Brock 1975; Newman 1997; Rabushka 2008).

In 1723, Pennsylvania and Delaware became the first colonies to initiate paper money systems that were not motivated by wartime fiscal crises. Their goal was to ameliorate internal economic crises caused by temporary depressions in their overseas trade balances. The paper money was to be removed from circulation by the end of the decade – presumably once the trade depression had passed. In 1729, paper-money advocates in Pennsylvania, such as Benjamin Franklin, won the debate to renew and expand the paper money experiment and turn it into a more or less permanent feature of the peacetime economy of these colonies (Grubb 2006). In 1733, Maryland became the third colony to initiate a paper money system that was not motivated by a wartime fiscal crisis. From the beginning, Maryland's paper money experiment was intended to be a permanent restructuring of the medium of exchange within the colony. Its goal was tied to transforming the transatlantic tobacco trade, which in turn required demonetarization of tobacco within the colony – Maryland's principal non-specie commodity money (Grubb 2008). The only colony issuing paper money subsequently to return to a specie standard for the remainder of the colonial period was Massachusetts in 1750, largely because of the rapid inflation that accompanied its paper emissions used to support military operations in King George's War (Officer 2005).

Following the English system, colonial paper currencies were denominated in pounds, shillings and pence (except for Maryland's money after 1766), but with the unit-of-account or proclamation exchange rate (par rate) to pounds sterling typically set higher than one-to-one. This par rate differed among the colonies, so for example there were 1.33 Maryland pounds and 1.67 Pennsylvania pounds to one pound sterling. Exchange rates to pounds sterling fluctuated considerably around, and sometimes departed from, these par rates. Some colonial legislatures made their paper money a legal tender for all transactions within their jurisdiction, some only for a subset of transactions, and some only for public debts (taxes). Some colonial currencies stated a specie exchange rate on their face, as did some early paper monies issued by New England colonies, Georgia and New Jersey, and Maryland paper money after 1766, but most did not (McCusker 1978; Newman 1997; Rabushka 2008). Colonial legislatures did not otherwise fix or defend an exchange rate between their paper money and specie coins. No colonial government succeeded in consistently exchanging its paper money on demand for specie coins in its colony, and most did not even try. Colonial treasuries did not typically keep specie reserves and so could not effectively act like banks. Colonial paper currencies seldom circulated far beyond the issuing colony's borders in any substantial quantities for any considerable period of time. Cross-colony and cross-oceanic trade was consummated using specie coins – the outside money – or credit in specie, often through bills of exchange.

As such, colonial paper monies are best thought of as inside monies on floating exchange rates to outside (specie) monies and to each other. They were true fiat currencies: that is, backed by nothing other than the promise that nominal taxes and mortgage payments owed to a government could be paid in the paper money issued by that government. Colonial legislatures passed taxes when issuing their paper monies that could be paid with these paper monies, or held mortgages on their subjects' lands in exchange for loaning them these monies. Often paper money emissions were redeemed via these tax and mortgage payments within a few years, with the money burnt upon redemption. This emission-redemption structure gave an immediate contemporaneous use and nominal anchor to colonial paper monies

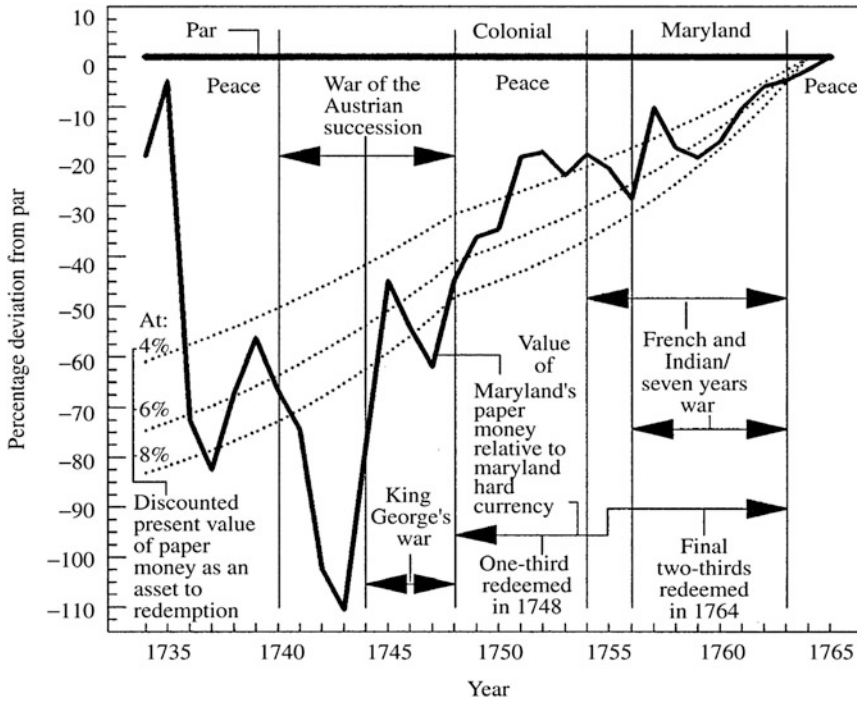
which supported their face value in current local exchange (Brock 1975; Rabushka 2008).

Several colonies, such as Rhode Island, New Jersey, Pennsylvania and Maryland, issued and redeemed portions of their paper money through land banks. Subjects borrowed paper money from their governments, pledging their lands as collateral. They could pay their mortgage principal and interest either in specie or in the paper money of their government, with the interest earned being an important source of income for some colonial governments. The amount any subject could borrow relative to the total sums available was typically restricted so that borrowings would be widespread (Rabushka 2008; Thayer 1953). In 1729, Benjamin Franklin argued that this land-bank method created a flexible money supply that passively expanded and contracted with the economy's money demand. This in turn produced a market-clearing monetary equilibrium that prevented excess quantities from being in circulation relative to demand and so prevented the paper money from depreciating. In essence, Franklin's argument was a primitive statement of the real bills doctrine, and may have been the first, and possibly only, statement of this idea by an American writer in the colonial era (Grubb 2006). Franklin may have taken the idea from John Law's 1705 Scottish land-bank pamphlet which was reprinted in London in 1720. Franklin visited London in the early 1720s. Whether colonial land banks actually functioned effectively as Franklin argued is yet to be conclusively determined.

The quantity of paper money in the initial authorization across colonies averaged 0.6 sterling-equivalents per capita, and ranged between 0.1 and 2.0. Thereafter, it stayed within this range, but averaged closer to 1.0. It only systematically exceeded 2.0 sterling-equivalents per capita in Rhode Island (1714–1746) and only briefly spiked above 2.0 in New Hampshire and New York during King George's War and the Seven Years' War, respectively. By contrast, the US money stock from 1795 to 1830 in sterling-equivalents per capita hovered between 1.4 and 2.2, with an average around 1.8 (Rousseau 2006).

The early experiments were often less than successful. The South Carolina pound in the late 1720s, the Maryland pound between 1736 and 1760, the Virginia pound between 1756 and 1765, and the Massachusetts pound in the 1740s suffered substantial depreciations. The British Parliament's response to the Massachusetts crisis was the Currency Act of 1751, which allowed colonies to issue paper money as long as it met two conditions: (1) that it not be a legal tender, and (2) that ample provisions (taxes) be put in place to redeem each issue within a reasonable time. While this Act applied only to New England, the Virginia crisis in the early 1760s led Parliament in 1764 to extend a version of the 1751 Act to all the colonies (Brock 1975; Ernst 1973; Rabushka 2008).

These early struggles were caused by excessive emissions relative to expected redemptions, which in turn were caused by perceived mismanagement in some cases and the overwhelming burden of war in other cases. The structure and backing of a paper emission could also affect its performance. For example, unlike other colonies, the 1733 Maryland paper pound was to be redeemed at par in specie by the Maryland government at designated future dates via a sinking fund. Most of the emission was handed out to its subjects in exchange for destroying trash tobaccos. Use of the money to pay contemporaneous local taxes was thwarted. Thus its value rested principally on the promised payoff in specie, one-third in 1748 and two-thirds in 1764. The colony taxed tobacco exports and invested the money in Bank of England stock at a rate that would generate the sums needed to meet the promised payoff, which the colony successfully executed. This structure, however, meant that the contemporaneous value of the Maryland pound relative to its par (face) value would track its present discounted value relative to its redemption date. In effect the Maryland pound was a zero-interest bearer bond (Grubb 2008; Rabushka 2008). Figure 1 shows this outcome, albeit with a lot of volatility around the discount trend early on (McCusker 1978). The Continental paper dollar issued by Congress during the American Revolution, of all colonial paper monies, most closely resembled the 1733–1764 Maryland pound. Before the Revolution, Benjamin Franklin had

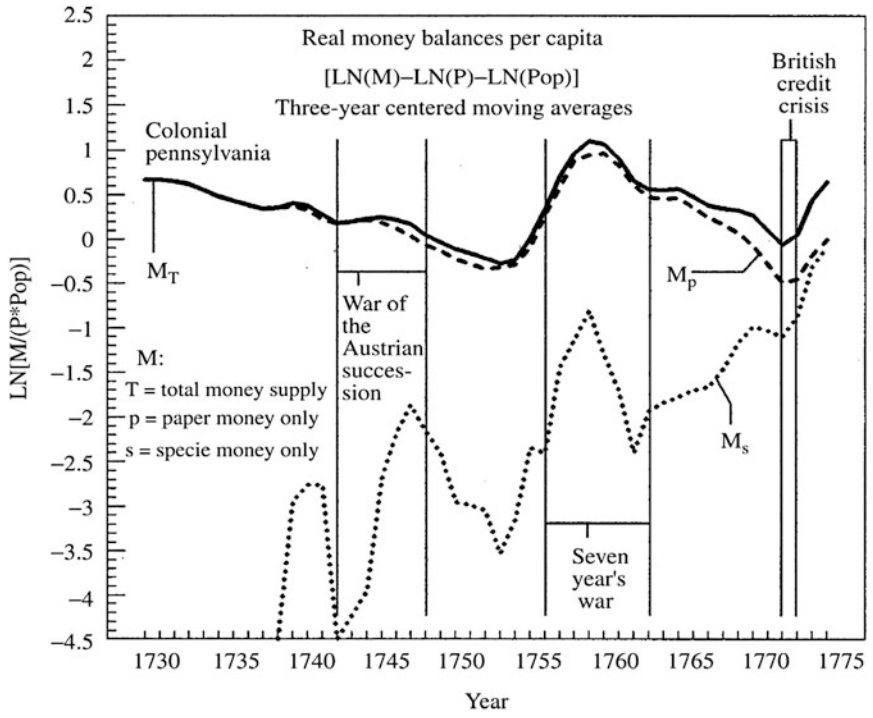


Money Supply in the American Colonies, Fig. 1 The value of Maryland's paper money, 1734–1765

noted the performance of the Maryland pound, as depicted in Fig. 1, with disapproval. While a general supporter of colonial paper money, his unexplained objection to the Continental dollar might be because its structure closely resembled that of the 1733–1764 Maryland pound (Grubb 2006). The idea, and occasionally practice, of having colonial (and later Continental) paper monies pay interest came from an effort to counterbalance the discount off their face value that resulted when they were backed by a bond-like redemption structure.

By 1750, 25 years before the Revolution, most colonies had learned how to maintain long-run price stability and manage their tax and mortgage-backed non-legal-tender inside paper money regimes (West 1978; Wicker 1985). For the most part, price indices in these colonies were trend-stationary between 1750 and 1775 with trends that did not exceed that experienced by colonies on a specie standard only (Grubb 2003). In addition, for the most part, each colony's paper money exchange rate to pounds sterling was stationary, and purchasing power parity cannot be rejected between each colony and between each colony and England. This performance is consistent with colonial legislatures successfully using their emission- redemption backing structures within a long-run quantity theory of money framework to manage their macro-economies ($MV = PY$ where M = money supply, V = velocity of money circulation, P = price level, Y = real output, and where the growth in V and Y are long-run constants). Substantial short-run volatility in velocity and real output per capita (which equals real money balances per capita) was still present (Rousseau 2007). For example, Fig. 2 shows the movement in real money balances per capita (ln(M/P*Pop) where Pop = population) for paper plus specie (total) money in Pennsylvania from 1729 to 1775. This series exhibits no trend and is stationary with a 3-year half-life to shocks (Grubb 2004).

The trade disruptions and wartime expenses of the American Revolution and its immediate aftermath (1775–86) strained these paper money systems. They often became associated with localized political trauma and economic chaos. Soon thereafter, the US Constitution, adopted by



Money Supply in the American Colonies, Fig. 2 The movement in per capita real money balances in Pennsylvania, 1729–1775

M

Congress in 1789, brought this colonial paper money system to an end by constitutionally barring national and state legislatures from issuing paper monies.

See Also

- Barter
- Commodity Money
- Fiat Money
- Inside and Outside Money
- Law, John (1671–1729)
- Money
- Quantity Theory of Money
- Purchasing Power Parity
- Real Bills Doctrine

Bibliography

Brock, L.V. 1975. *The currency of the American colonies, 1700–1764*. New York: Arno.

Ernst, J.E. 1973. *Money and politics in America, 1755–1775*. Chapel Hill: University of North Carolina Press.

Grubb, F. 2003. Creating the U.S.-dollar currency union, 1748–1811: A quest for monetary stability or a usurpation of state sovereignty for personal gain? *American Economic Review* 93: 1778–1798.

Grubb, F. 2004. The circulating medium of exchange in colonial Pennsylvania, 1729–1775: New estimates of monetary composition, performance, and economic growth. *Explorations in Economic History* 41: 329–360.

Grubb, F. 2006. *Benjamin Franklin and the birth of a paper money economy.* Philadelphia: Federal Reserve Bank of Philadelphia. Available at: http://www.philadelphiafed.org/education/index.html#publications. Accessed 8 Nov 2008

Grubb, F. 2008. *Creating Maryland's paper money economy, 1720–1739: The role of power, print, and markets*. NBER working paper #13974. Available at: http://www.nber.org/papers/w13974. Accessed 8 Nov 2008.

McCusker, J.J. 1978. *Money and exchange in Europe and America, 1600–1775*. Chapel Hill: University of North Carolina Press.

Mossman, P.L. 1992. *Money of the American colonies and confederation: A numismatic, economic and historical correlation*. New York: American Numismatic Society.

Newman, E.P. 1997. *The early paper money of America*, 4th ed. Iola: Krause.
Officer, L.H. 2005. The quantity theory in New England, 1703–1749: New data to analyze an old question. *Explorations in Economic History* 42: 101–121.
Rabushka, A. 2008. *Taxation in colonial America*. Princeton: Princeton University Press.
Rousseau, P.L. 2006. A common currency: Early US monetary policy and the transition to the dollar. *Financial History Review* 13: 97–122.
Rousseau, P.L. 2007. Backing, the quantity theory, and the transition to the US dollar, 1723–1850. *American Economic Review–Papers and Proceedings* 97: 266–270.
Thayer, T. 1953. The land-bank system in the American colonies. *Journal of Economic History* 13: 145–159.
West, R.C. 1978. Money in the colonial American economy. *Economic Inquiry* 16: 1–15.
Wicker, E. 1985. Colonial monetary standards contrasted: Evidence from the Seven Years' War. *Journal of Economic History* 45: 869–884.

Money, Classical Theory of

Roy Green

Abstract

An integral part of the classical theory of value and distribution, the classical theory of money emerged largely in response to the issue of the relationship between changes in the money supply and the price level. This issue was central to the Price Revolution of the 16th and 17th centuries, the Napoleonic war inflation and the industrial crises of the mid-19th century. It was not the existence of an empirical correlation that was in dispute, but the direction of causation. A solution would therefore require a *theoretical* approach as well as knowledge of the facts.

Keywords

Bank Charter Act 1844 (UK); Bank Restriction period; Banking School; Bullion Report (1810) (UK); Bullionist controversies; Cantillon, R.; Classical dichotomy; Classical law of circulation; Classical theories of distribution; Classical theory of money; Convertibility; Credit; Currency School; Equation of exchange; Fiduciary money; Hume, D.; Inflation; Law of reflux; Law, J.; Metallic money; Monetarism; Money, classical theory of; Money supply; Paper money; Price Revolution; Quantity theory of money; Real bills doctrine; Say's Law; Specie-flow mechanism; Steuart, J.; Tooke, T.; Torrens, R.; Value theory

JEL Classifications

E4

The classical theory of money is an integral part of the classical theory of value and distribution; and its conceptual categories have real counterparts in historical experience. These categories begin with metallic money and progress to the more complex forms of fiduciary money and credit.

Classical Framework

The equation of exchange forms a common point of reference for all approaches to monetary theory, since the relationships it expresses simply constitute a truism and do not in themselves imply causality: $MV = PT$, where M denotes the money supply, V the velocity of circulation, P an index of prices and T the number of commodity transactions. This equation may also be written: $MV = PY$, where Y denotes total output, the index P is correspondingly adjusted and V no longer reflects the circulation of a stock of commodities but the rate of expenditure of a flow of income (corresponding to the flow of output). We use this alternative formulation to specify the classical approach to monetary theory. The only difference of substance is the replacement of the sum of commodity transactions with a measure of net output over a given period, hence excluding non-produced assets (such as land) from the exchange process.

The classical theory of money was developed largely as a response to the practical issue of the relationship between changes in the money supply and the price level. This issue was central to three historical episodes which form the background to

our discussion: the Price Revolution of the 16th and 17th centuries, the Napoleonic war inflation and the industrial crises of the mid-19th century. It was not the existence of an empirical correlation that was in dispute, but the direction of causation. A solution would therefore require a *theoretical* approach as well as knowledge of the facts.

The basic structure of the solution arose from discussion of the Price Revolution. Instead of augmenting wealth in the manner suggested by mercantilist doctrine, the influx of gold and silver from the newly discovered American mines seemed only to devalue the unit of account. An immediate interpretation was offered by the quantity theory of money, which attributed the increase in the price level throughout Europe entirely to monetary expansion. According to David Hume, money had no intrinsic value and was simply a means of circulation, in which capacity it served simultaneously as money of account (1752, p. 33). This approach 'essentially amounted to treating money not as a commodity but as a voucher for buying goods' (Schumpeter 1954, p. 313). Once in circulation, money acquired merely a 'fictitious value', whose magnitude was established by demand and supply (Hume 1752, p. 48; also Montesquieu 1748, pp. 50–1; Vanderlint 1734, pp. 2–3; Locke 1691, p. 233).

Classical economists, by contrast, treated money as a *real commodity*, whose value was determined like other commodities by the labour time socially necessary for its production (Petty 1963, vol. 1, pp. 43–4; Smith 1776, p. 24; Ricardo 1821, pp. 85–6). They traced the cause of the Price Revolution not to monetary phenomena but to lowered production costs at the mines (Nef 1941; Outhwaite 1969, esp. p. 29; Vilar 1976, esp. p. 343). It followed that, in the long run, when economic activity is regulated by permanent forces, the magnitude of P in the equation of exchange is determined on the basis of value theory and both Y and V are fixed due to Say's Law and institutional factors respectively. Hence P is the independent variable in the equation and M the dependent variable. Any movement in P as a result of changes in the production costs of commodities (or money) has a commensurate effect on M. This determination of aggregate monetary requirements in the 'real' sector of the economy became known as the 'classical dichotomy' and constitutes the basic classical law of circulation (Petty 1963, vol. 1, p. 36; Smith 1776, pp. 332–3; Ricardo 1923, p. 158; Marx 1867, pp. 123–4). In other words, causation runs from prices to money in classical economics and not the reverse as we find in both traditional quantity theory and neoclassical monetarism (Eatwell 1983; Green 1982). All things being equal, 'The quantity of money that can be employed in a country must depend on its value' (Ricardo 1821, p. 352). The type of money employed in the circulation process has no bearing on this conclusion, since V will be determinate whatever *its* numerical value.

Had the scope of classical economics extended no further than the study of permanent economic forces, the question of whether it possessed a 'quantity theory of money' would not have arisen. But the limitations of a long-run approach in explaining concrete developments and formulating relevant policies convinced most classical writers to take into account the role of temporary factors. In particular, the effect of exogenous changes in the money supply needed to be explained. Now the problem became complicated by the definition of money and the nature of financial organization. If Say's Law kept Y constant, only two possibilities remained open: a price adjustment, that is, a change in P, or a quantity adjustment, that is, a change in V (by hoarding or dishoarding). This was the essence of the division among the classical economists. One group was led by Ricardo and included the bullionists (that is, supporters of the 1810 Bullion Report), and later, the Currency School. The other group comprised the anti-bullionists and the Banking School and was given qualified approval by Marx.

The dominant Ricardian group held consistently that both Y and V were always fixed. The quantity 'theory' of money was therefore no theory at all in this view, but simply a logical outcome of assuming Say's Law. The inflationary process was seen as the transitional mechanism by which monetary deviations were corrected: 'That commodities would rise or fall in price, in proportion to the increase of diminution of money,

I assume as a fact which is incontrovertible' (Ricardo 1923, p. 93 fn., emphasis added).

The opponents of quantity theory, on the other hand, were prepared to sacrifice logical consistency in an attempt to interpret the real events with which they were confronted. Their often pioneering expositions generally placed the weight of adjustment on *V*, although the extent was seen as contingent upon the composition of *M* – whether the money supply was metallic, fiduciary or credit. The flaw in their approach was their failure to overthrow Say's Law and develop an analysis of the saving–investment process, that is, a theory of output. Had they done so, their challenge to the incorporation of quantity theory into classical economics may have been more successful.

Currency and Credit

By the time the Bank of England suspended cash payments in 1797, a body of principles on the role and behaviour of paper money had already been formed. The collapse of Law's system led to considerable discussion which culminated in Smith's authoritative exposition of banking in the *Wealth of Nations*. There Cantillon's view was accepted – as against Law and Steuart – that banking could not increase the quantity of capital but only its turnover (Smith 1776, p. 246). This accorded with the given output assumption of Say's Law. It was also established that paper money would not depreciate provided its total amount did not exceed the value of gold and silver that would otherwise have circulated at any given level of economic activity (1776, p. 227).

More contentiously, Smith argued that the economic convertibility of paper and metallic money could be maintained not only by enforcing legal convertibility but also by having banks adopt the practice of discounting 'real bills', that is, securities backed by real assets (1776, p. 239 and *passim*). This became known as the 'real bills doctrine'. It was repudiated first by Thornton and then by Ricardo and the Currency School, but rehabilitated as the 'law of reflux' by the Banking School.

The Bank Restriction period was marked by high inflation accompanied by a rise in the market price of bullion over its mint price. This indicated a depreciation of paper currency in terms of the monetary standard, a phenomenon which could not have existed when convertibility was enforced by law. The central problem was to explain the appearance of this premium on bullion, and to find a principle whose practical implementation would restore and maintain economic convertibility, thus ensuring that the bank notes conformed to the behaviour of metallic currency. The explanation which gained widest acceptance was based upon the quantity theory of money. It was presented officially in the Bullion Report and then developed by Ricardo. The remedy for inflation implied by this approach was control over the money supply by the authorities.

Ricardo began his analysis by recognizing the need to replace gold and silver in the sphere of circulation by paper – provided only that it was issued in the same amount, that is, the amount prescribed by the value of the metal which served as the monetary standard: 'A currency is in its most perfect state when it consists wholly of paper money, but of paper money with an equal value with the gold which it professes to represent' (Ricardo 1821, p. 355). Ricardo's discussion of legally convertible bank notes followed Smith, with some of Thornton's modifications. Since their equivalence with gold was guaranteed, they could not be issued in a greater quantity than the value of the coin which would otherwise have circulated. Any attempt to exceed this sum would precipitate a return of notes for specie, a depreciation of both paper and metallic currency, and the subsequent export of superfluous bullion (Ricardo 1923, pp. 7–13). Overextension of inconvertible notes in a 'mixed currency' of notes and coins had the same effect so long as the degree of excess was no greater than the amount of coin in circulation (1923, p. 13, n., pp. 108–12).

In 1809, however, when Ricardo entered the bullion controversy, the currency was composed almost entirely of inconvertible paper. He therefore ascribed the rise in commodity prices, in so far as it corresponded with the premium on bullion, wholly to monetary overissue. Such an

overissue would have no other effect than to 'raise the *money* price of bullion without lowering its *value*, in the same manner, and in the same proportion, as it will raise the prices of other commodities'. In other words, although paper money was depreciated, the 'bullion price' of commodities was unaltered. Hence the deterioration of the foreign exchanges 'will only be a *nominal*, not a *real* fall, and will not occasion the exportation of bullion' (1923, p. 13 n. and p. 109).

Ricardo was criticized for ignoring the real reasons for the inflation, which had more to do with harvest failures, war subsidies and the Napoleonic blockade (Morgan 1965, pp. 46–7). Moreover, he left himself open to the charge of superimposing a theory of *fiduciary* money on a *credit* system. Had bank notes been issued at will by the state, Ricardo would have been correct in his characterization of their relationship to the price level. Fiduciary money only *represents* gold in the circulation process, and is depreciated to the extent of its overissue. The depreciation persists until the quantity is reduced, for there are no self-correcting tendencies as in the case of convertible paper. However, the fact that the notes of the Bank Restriction period were not forced currency but credit responding to the demand of the non-bank public was excluded from Ricardo's consideration by Say's Law. He treated the notes as though they were fiduciary because output and velocity were independently given. The possibility of disintermediation when the authorities tried to contract the note issue was also excluded. The fixed velocity assumption implied that the rest of the spectrum of credit would shrink commensurately with the notes. In fact, as the Banking School was to demonstrate, credit instruments simply expanded in their place.

The resumption of specie payments in 1819 on the advice of Ricardo and the bullionist spokesmen did nothing to eliminate price instability from Britain's developing industrial economy. In 1825 and 1836, phases of vigorous expansion ended with an adverse balance of payments, a gold drain from the Bank of England and an inflationary collapse into recession. The Currency School – a new orthodoxy which Morgan describes as the 'heirs of the Bullion Report' – attributed the recurrent dislocation to excessive monetary growth. The convertibility of bank notes was no longer seen as a sufficient safeguard against over issue and consequent depreciation. The Currency School argued that rules would have to be devised to make the paper currency fluctuate as though it were metallic, in other words to replicate the 'automatic' operation of Ricardo's international specie-flow mechanism. This implied regulation of the note issue by the monetary authorities in strict conformity with the foreign exchanges; the export and import of bullion was treated as an index of monetary excess or deficiency, and thus of the value of the notes.

The currency principle was given practical effect by the Bank Charter Act of 1844, which set the pattern of the UK financial system for almost a century. It was challenged by the Banking School, which Morgan calls 'the heirs to the opposition to the Bullion Report, but the opposition as it might have been rather than as it was'.

The long-run determination of aggregate monetary requirements by nominal output – the 'supply side' of the equation of exchange – was common ground in the debate. The real point at issue was again the *short-run* behaviour of the variables. Whereas the Currency School adopted Ricardian quantity theory and applied it to a credit system made up of convertible bank notes, the Banking School took the alternative view of metallic circulation and tried to develop a theory specific to credit. Both sides recognized the importance of theorizing the laws of metallic circulation as a precondition for the analysis of paper currency. The entire Currency School case for monetary control rested upon the assertion that the note issue would not by itself emulate the behaviour of a metallic system. Despite legal convertibility, it might depart at least temporarily from the amount and value of the metallic money which would otherwise have circulated. In practice, therefore, economic convertibility could be ensured only by quantitative intervention on the part of the authorities (Torrens to Lord Melbourne, *cit*. Tooke 1844, p. 7).

Banking School criticism took three main lines. First, starting from the assumption that legal convertibility necessarily implied economic

convertibility, they pointed out that any discrepancy between the note issue and a purely metallic system arose from the Currency School's erroneous theory of metallic circulation rather than from the supposed autonomy of the notes. Second, any effect of prices attributed to bank notes could not be denied to a range of financial assets excluded by the Currency School from their definition of money. Third, bank notes were in any case not money but credit, and therefore never could be over issued, through the credit structure as a whole might be extended beyond the limits of real accumulation by 'speculation and overtrading'.

The Banking School emphasized that the volume of notes in circulation could not be increased at will by the authorities, but only in response to the demand of the non-bank public. This crucial difference between fiduciary money and bank notes was explained by Tooke as consisting, 'not only in the limit prescribed by their convertibility to the amount of them, but in the *mode of issue*' (Tooke 1844, pp. 70–1, emphasis added; see also Fullarton 1845, ch. 3, and Wilson 1859, pp. 48, 51–2, 57–8). The currency principle, by contrast, 'completely identifies *monetary turnover* with *credit*, which is economically wrong' (Marx 1973, p. 123). An advance of bank notes did not *add* to the money supply, but merely changed its *composition*, allowing the substitution of one financial asset for another in the hands of the public. Excess notes returned automatically to the bank 'in the shape of deposits or of a demand for bullion' (Tooke 1844, p. 60; see also Wilson 1859, p. 58; Marx 1867, III, pt. 5). This was the basis of the law of reflux, which Fullarton called 'the great regulating principle of the internal currency'. It held that economic convertibility could be ensured not only by a legal right to exchange notes for specie but also by maintaining a balance between the notes advanced on loan and those returned to the bank at maturity. Provided lending took place on commercial paper which represented a real or (within a given timescale) potential sum of values, 'the reflux and the issue will, in the long run, always balance each other' (1845, pp. 64–7; also p. 207; Marx 1973, p. 131).

The Banking School did not imagine that the economic cycle could be eliminated by monetary measures. Instead, they evolved a new set of criteria by which the authorities could operate on the 'state of credit' through interest rate and reserve management (Tooke 1844, p. 124; Fullarton 1845, p. 164; Marx 1867, III, 447). In practice, all that lay between the Currency and Banking Schools was ultimately a matter of timing, but this reflected profound theoretical differences. Within the framework of classical analysis, it was the Banking School which came closer to constructing a modern philosophy of monetary regulation.

See Also

- ▶ Banking School, Currency School, Free Banking School
- ▶ Bullionist Controversies (Empirical Evidence)
- ▶ Hume, David (1711–1776)
- ▶ Quantity Theory of Money
- ▶ Ricardo, David (1772–1823)
- ▶ Thornton, Henry (1760–1815)

Bibliography

Cantillon, R. 1775. *Essai sur la nature du commerce en général*. London: Macmillan, 1931.

Eatwell, J. 1983. The analytical foundations of monetarism. In *Keynes's economics and the theory of value and distribution*, ed. J. Eatwell and M. Milgate. London: Duckworth.

Fullarton, J. 1845. *On the regulation of the currency*. London: John Murray.

Green, R. 1982. Money, output and inflation in classical economics. *Constitutional Political Economy* 1: 59–85.

Hume, D. 1752. *Writings on economics*. Oxford: Oxford University Press, 1955.

Law, J. 1705. *Money and trade considered*. Edinburgh: Anderson.

Locke, J. 1691. Consequences of the lowering of interest and raising the value of money. In *Principles of political economy*, ed. J.R. McCulloch. London: Ward, Lock & Co., 1825.

Marx, K. 1867. *Capital*. Moscow: Progress Publishers, 1971.

Marx, K. 1973. *Grundrisse*. Harmondsworth: Penguin.

Montesquieu, C. 1748. *The spirit of laws*. London: George Bell & Sons, 1900.

Morgan, E.V. 1965. *The theory and practice of central banking, 1797–1913*. London: Frank Cass.

Nef, J.U. 1941. Silver production in central Europe: 1450–1618. *Journal of Political Economy* 49: 575–591.
Outhwaite, R.B. 1969. *Inflation in Tudor and early Stuart England*. London: Macmillan.
Petty, W. 1963. *The economic writings of Sir William Petty.* New York: Kelley.
Ricardo, D. 1821. *Principles of political economy and taxation* (Ed. P. Sraffa). Cambridge: Cambridge University Press, 1951.
Ricardo, D. 1923. *Economic essays*. London: Frank Cass.
Schumpeter, J. 1954. *A history of economic analysis*. London: Allen & Unwin.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: Routledge, 1890.
Steuart, J. 1767. *An inquiry into the principles of political economy.* Edinburgh: Oliver and Boyd, 1966.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*. London: LSE reprint series, 1939.
Tooke, T. 1844. *An inquiry into the currency principle*. London: LSE reprint series, 1959.
Vanderlint, J. 1734. *Money answers all things*. London: T. Cox.
Vilar, P. 1976. *A history of gold and money: 1450–1920*. London: New Left Books.
Viner, J. 1937. *Studies in the theory of international trade*. London: Allen & Unwin.
Wilson, J. 1859. *Capital, currency and banking*, 2nd ed. London: The Economist.

Moneylenders

Amit Bhaduri

The standard practice of moneylending in the unorganized credit market differs from financial intermediation by the commercial banks. Banks operating on the basis of a 'fractional reserve system' hold in cash reserve only a fraction of their total debt obligation to the public. In effect, this becomes the method of creating credit-money by the banks through the so-called 'credit multiplier'. However, there exists no ready counterpart to such credit creation by private moneylenders in the unorganized markets. In principle, a private moneylender with a good reputation for solvency can also create *private* debt obligations in the form of personal promisory notes or I-owe-you's (IOUS). And, the issue of such *private* debt obligations can even be several times the cash in reserve with him, in analogy with the credit multiplier of commercial banking. Although such *private* debt obligations are not uncommon in some less monetized rural areas or in the informal banking sector (coexisting side by side with the formal banks in urban centres) in many underdeveloped countries, the issuing of such private debt obligations must be intrinsically far more restricted in scope for at least three reasons. First, without either a legally stipulated 'cash reserve ratio' or an institutional 'lender of the last resort', private moneylenders have to rely entirely on their personal creditworthiness, in case there develops a sudden 'run' on their debt obligations. Second, *personal* reputation must normally be *spatially* restricted to relatively small areas. In turn, this tends to fragment the unorganized market for credit. Finally, many private moneylenders, especially in the poorer rural areas of underdeveloped countries do *not* act as proper financial *intermediaries*. Instead, as the name suggests, they are primarily lenders of money (usually out of their own savings), but not takers of deposits. And, not being financial intermediaries, the income or profit for this class of moneylenders cannot be explained in terms of the margin of their lending rate over the deposit rate, unlike in the case of commercial banks. This suggests a different mode of operation in terms of the profitability of private moneylending in the unorganized market. For example, the private moneylender must get a *higher* rate of return on the loan he advances from his personal savings than he could secure from deposits with banks in the organized credit market to make such activities economically worthwhile for him. It is the task of economic theory to explain how this may come about *without* financial intermediation.

Empirical studies abound (see Bhaduri 1983, ch. 5 for references to Indian field studies; Nisbet 1967, for Chile; and Tun Wai 1957–8 for some earlier evidence) to suggest that in many underdeveloped countries, the rate of return on private moneylending is indeed considerably higher than say, the deposit or lending rate offered by banks. Although some economists (e.g. Bottomley 1975)

have tried to explain this in terms of the *lender's* risk margin, any such explanation can be seen to be inadequate from our preceding discussion. The lender's risk margin is supposed to cover the loss of the defaulted fraction of principal lent. According to this view a typical moneylender expects a certain proportion (q) of the total loan he advances to be defaulted and charges a sufficiently high rate of interest (i) on the loan expected to be paid back to him ($1 - q$) to cover his capital loss. In this case, his overall rate of return (r) is more or less the same as the rate of deposit (or the lending rate) with the banks in the organized credit market. This means,

$$(1-q)(1+i) = (1+r) \text{ or, } i = r + q/(1-q).$$

Clearly, given $1 > q > 0$ the private money lender's lending rate i would exceed that in the organized market (r) as it also covers the lender's risk margin due to default.

Such a theory starts with the presumption that the rate of return is 'competitively' equalized between the organized and the unorganized credit market (e.g. at 'r' in the above calculation). It is argued therefore, that even when private moneylenders do *not* operate as financial intermediaries, they in effect earn more or less the same rate of return as the typical financial intermediaries in the organized sector. The assumption underlying is that, the organized and the unorganized credit markets are thoroughly integrated. In a similar manner, explanations of higher interest rates on private lending through higher 'administrative cost' of managing such loans or higher transaction cost in general adds analytically nothing new: it is always viewed as a lender's margin over the interest rate in the organized market. Thus, explicitly or implicitly this view relies on the assumption that the organized and unorganized credit markets are integrated.

Empirically, the integration of the organized and the unorganized credit market is open to serious doubt. This is most strikingly brought out by the nearly universal fact that the poorest strata of the peasantry in many underdeveloped countries rely heavily, if not exclusively, on private moneylenders and, *not* on sources of institutional finance. Indeed, financial institutions like banks and credit cooperatives typically do not consider them creditworthy. But paradoxically enough private moneylenders do consider them creditworthy for advancing loans (Bhaduri 1983, pp. 12–16). This would suggest *lack* of integration between the organized and the unorganized credit market, at least insofar as the criteria for creditworthiness are concerned.

Creditworthiness of a borrower generally (i.e. 'reputation' apart) depends on the collateral securities that he can offer against the loan advanced to him. The typical collateral securities that a very poor peasant can offer, for example already encumbered land, standing crops of his future labour service, do not usually have well-defined market prices. Consequently, they are not acceptable as collateral securities to usual institutional lenders like banks. On the other hand, a *local* village moneylender is willing to accept them as collateral because, either he can make personal use of them (e.g. an agriculturist moneylender may be happy to obtain the use-right of an already encumbered piece of land or the future labour service of a defaulted borrower); or, he can undervalue such collaterals substantially in a loan arrangement. In the latter case, he would indeed make some 'capital gains' in case of default of loan by the borrower, as the undervalued asset gets transferred to him in case of default. This suggests a basic difference in the mode of operation of the organized and the unorganized, private credit market. In the organized market, there is risk of capital loss to the lender in case of default. However, in the unorganized market, this risk may be largely avoidable by the lender when he is in a position to sufficiently undervalue the collaterals. Indeed, default would then mean capital gain rather than capital loss to him through the transfer of such undervalued collaterals. And, because the borrower is threatened with possible capital loss in case of default, such loan arrangements can be seen to be characterized by *borrower's rather than lender's risk* (Bhaduri 1983, ch. 5). Further, in such a loan arrangement, the lender would have

a tendency to *induce* default by charging exceptionally high interest rates in order to make capital gains. This can appropriately be described as the method of *usury,* when default induced through high interest charge results in capital gains to the lender. Indeed, most 'pawn-shops' are also known to operate on a similar principle.

The helpless borrower usually goes to the pawn-shop or to the private moneylender in rural areas because he is not considered sufficiently creditworthy to obtain loans in the organized credit market. It should be evident that his helplessness as a borrower is more acute, the more desperately he needs the loan (e.g. consumption loan for survival). Analytically, the more *inelastic* is the borrower's demand function for loan, the more vulnerable he would be to this method of usury described earlier. As a matter of fact, his only defence in the extreme case of totally inelastic demand for loan may simply lie in *deliberately* defaulting, if the interest rate is raised too high by the lender. In that case, he accepts to lose his collateral asset instead of trying to meet the high interest charge. However, when his loan demand function is more elastic – for example, he can decide to borrow less if the interest rate is pushed higher or the price of the collateral is pushed lower – the borrower is placed more favourably in terms of bargaining power. In such cases, the lender would *simultaneously* decide what interest rate to charge and what collateral price to offer (Basu 1984). It is conceivable that the lender would charge a lower interest rate to entice the borrower to take a larger amount of loan; but at the same time, he would also undervalue collaterals in the hope that the borrower would not be able to pay back that larger loan so that the lender would again make capital gains through asset transfers.

See Also

- ▶ Agriculture and Economic Development
- ▶ Fiscal and Monetary Policies in Developing Countries
- ▶ Peasant Economy
- ▶ Sharecropping

Bibliography

Tun Wai, U. 1957–8. Interest rates outside the organized money markets of underdeveloped countries. *IMF Staff Papers* 6: 80–142.

Moneylenders in Developing Countries

Hanan G. Jacoby

Abstract

Moneylenders are a principal source of credit in developing countries. They thrive where collateral is scarce and legal enforcement of debt contracts is difficult. Their advantages over banks include better knowledge about creditworthiness of their clientele and greater ability to enforce repayment. Landlords lend to their sharetenants because they can capture a larger share of the tenants' surplus than can outside lenders. Other credit by moneylenders is in kind, such as in the form of input advances or deferred rent. The effects of government credit policies depend importantly on the relationship between moneylenders and banks.

Keywords

Agricultural finance; Asymmetric information; Credit markets in developing countries; Informal sector; Moneylenders in developing countries

JEL Classifications

O1

Moneylenders are a principal source of credit in developing countries, especially in rural areas, but are notoriously difficult to classify. They may be shopkeepers, millers, traders, landlords, or professional financiers. Moneylenders operate within a broad spectrum of lending 'formality' bounded

above by the activities of commercial or agricultural banks and below by credit from friends, relatives, and fellow clan members. Banks normally take deposits, ask lenders for collateral, have formal procedures for loan applications with written contracts, and operate within the legal system; moneylenders may do none of the above. Friends, relatives and clanmembers, on the other hand, do not require their loans to be secured, make verbal agreements, generally do not charge interest, and often allow state-contingent repayment (Udry 1994). Reciprocity and social pressure, rather than legal sanctions, enforce such kin- or clan-based credit (La Ferrara 2003). Moneylenders, by contrast, are less flexible about the terms of repayment, more likely to charge interest, and less able to mobilize social opprobrium to punish default.

Formal sector lending is limited by the value of collateral, which in agricultural areas is usually in the form of land. Land is useful as collateral only to the extent that it can be legally repossessed upon default of the loan. This, in turn, requires that land be titled, or that ownership be otherwise documented, and that foreclosure be enforceable in court. Moneylenders thrive in settings where collateral is scarce or legal enforcement of debt contracts is weak or non-existent. But such conditions are not sufficient for the presence of moneylenders, who ultimately face the same problem as do banks; earning profit in the face of potential default. One way to do so is by setting a low interest rate and rationing credit, as in Stiglitz and Weiss (1981). This presumes, however, that moneylenders have no particular informational advantage over banks.

What, then, is the comparative advantage of the moneylender? There are three, not mutually exclusive, answers to this question, all related to the fact that the moneylender either resides in the same village or locality as his clientele, and is thus likely to have much more personal knowledge of and contact with them than would a bank, or is simultaneously dealing with his borrowers in another market. By virtue of proximity, a moneylender may, first of all, have a better idea as to whether a borrower can successfully implement a given project and thus repay the loan. In other words, it is mainly the bank, not the moneylender, that faces asymmetric information about the creditworthiness of the borrower.

A second advantage the moneylender may have over a bank is in enforcing repayment. Traders or millers often advance credit against the forthcoming harvest. By acquiring the right to market his debtor's harvest as a condition of the loan, and to deduct principal and interest at the time of sale, the trader–lender effectively guarantees debt seniority. Indeed, the trade–credit linkage may serve the dual purpose of enforcement and screening. The frequent exclusivity of such marketing agreements insures that the lender can observe the *entire* output of his borrowers, so as to monitor their ability to repay, as well as that of his prospective borrowers, so as to assess their future creditworthiness; at the same time, no other lender can have access to this information and thereby compete away borrowers (Siamwalla et al. 1990; Aleem 1990). Moneylenders may also have more effective means of preventing their clients from absconding with the loan principal or diverting it to non-productive uses (Giné 2005). While banks cannot legally prevent such strategic default beyond confiscating what collateral they hold, moneylenders may be able to exert various kinds of physical and psychological pressures to ensure repayment.

Lastly, moneylenders may more readily exchange information about borrowers' repayment histories than banks in developing countries. An informal borrower with a reputation for default will not only be unable to obtain future loans from the same moneylender but may lose access to all local creditors. Kletzer and Wright (2000) show that, when credit histories are public information, punishing default by a debt moratorium until such time as the lender is repaid is a credible strategy. If any competing moneylender fails to respect this punishment by subsequently lending to the delinquent borrower, the other moneylenders can induce the borrower to default on this loan by offering him a better deal, thus 'cheating the cheater'. When there is a high enough probability that credit histories are 'forgotten' or hidden, however, this type of equilibrium breaks down (Hoff and Stiglitz 1997). Reputation equilibria are thus sensitive to the extent of village information networks, about which little is known.

These arguments aside, collecting on a past debt may not be an unalloyed benefit to the moneylender. When the borrower's output or investment depends importantly on his unverifiable effort, debt creates an incentive problem. The higher the debt burden, the more the borrower is working merely to pay off the loan, the less willing he is to work, the lower his output, and, consequently, the more likely he is to default. Given 'debt overhang', the moneylender has to trade off higher debt collection against higher probability of default and collecting nothing. The resolution may involve forgiving debt. Evidence on the extent of debt forgiveness in informal credit markets is lacking (Fafchamps and Gubert 2004, is a notable exception), but there are at least two reasons to believe that it is not widespread. First, in a long-term credit relationship, the moneylender has the option of rescheduling debt in the hopes that the borrower's fortunes will improve, a less drastic step than forgiveness. Second, the impact of forgiveness on incentives is diluted when the borrower and lender are not in an exclusive credit relationship. Since other creditors can free ride on the lowering of total debt, there may be too little forgiveness in equilibrium.

Landlord–moneylenders have motivated a considerable literature on 'interlinked' tenancy-credit contracts. Because the landlord must always verify the harvest of a share-tenant, he is in a better position to enforce debt repayment than an outside moneylender. Perhaps more importantly, however, the landlord has a stronger incentive to provide credit to his share-tenant than any other moneylender. This is because the landlord, in general, captures a larger share of incremental surplus due to an increase in the tenant's working capital than does an otherwise equivalent outside moneylender (see, for example, Basu et al. 2000). Even if the landlord himself faces relatively high credit costs, given his enforcement advantage, he may still prefer to on-lend funds to his tenant from a moneylender under a so-called 'credit-layering' arrangement (Mansuri 2007).

The boundaries of moneylending are further obscured by the multifarious nature of credit. Traders, for example, often advance inputs in kind rather than cash, with interest collected through a markup on the price. Burkhart and Ellingsen (2004) rationalize this form of trade credit on the grounds that inputs are less easily diverted to non-productive uses than cash; in-kind loans thus alleviate a monitoring problem. Another form of in-kind lending occurs when landlords defer rental payments until after the harvest. Besides the possible monitoring advantage, such debt contracts have better incentive properties than share-contracts when the tenant's liability is limited (Innes 1990) or when tenant risk aversion and yield variability are not too high (Arimoto 2005). Since land is far and away the most important factor of agricultural production, the value of deferred rent may dwarf that of other seasonal borrowing.

Interest in moneylenders has centred around their role in modulating the impact of government policies, such as interest rate subsidies or controls, that can be effectively implemented only in the formal sector. The effects of such policies depend critically on the relationship between moneylenders and banks. The literature has taken two approaches to the formal–informal sector interaction. The first assumes a vertical structure whereby moneylenders act as middlemen, borrowing from the formal sector and on-lending to uncollateralized peasants (Hoff and Stiglitz 1997; Floro and Ray 1998). In the second approach, moneylenders and bankers compete with one another, with the residual demand for credit in the formal sector spilling over into the informal sector. Bell et al. (1997) and Kochar (1997) have moneylenders coexisting with banks by virtue of exogenous ceilings on formal sector credit, whereas Giné (2005) and Jain (1999) explicitly model moneylenders' informational advantage over banks to obtain coexistence in equilibrium without imposing formal sector credit rationing.

See Also

- ▶ Adverse Selection
- ▶ Agricultural Finance
- ▶ Microcredit
- ▶ Sharecropping

Bibliography

Aleem, I. 1990. Imperfect information, screening and the costs of informal lending: A study of rural credit markets in Pakistan. *World Bank Economic Review* 4: 329–349.

Arimoto, Y. 2005. State-contingent rent reduction and tenancy contract choice. *Journal of Development Economics* 76: 355–375.

Basu, K., C. Bell, and P. Bose. 2000. Interlinkage, limited liability and strategic interaction. *Journal of Economic Behavior and Organization* 42: 445–462.

Bell, C., T. Srinivasan, and C. Udry. 1997. Rationing, spillover and interlinking in credit markets: the case of rural Punjab. *Oxford Economic Papers* 49: 557–587.

Burkhart, M., and T. Ellingsen. 2004. In-kind finance: a theory of trade credit. *American Economic Review* 94: 569–590.

Fafchamps, M., and Gubert, F. 2004. Contingent loan repayment in the Philippines. Discussion paper no. 215. Department of Economics, Oxford University. Online.

Floro, M., and D. Ray. 1998. Vertical links between formal and informal financial institutions. *Review of Development Economics* 1: 34–56.

Giné, X. 2005. *Access to capital in rural Thailand: An estimated model of formal vs. informal credit*. Policy Research working paper no. 3502. Washington, DC: World Bank.

Hoff, K., and J. Stiglitz. 1997. Moneylenders and bankers, price–increasing subsidies in a monopolistically competitive market. *Journal of Development Economics* 52: 429–462.

Innes, R. 1990. Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory* 52: 45–67.

Jain, S. 1999. Symbiosis vs. crowding-out, the interaction of formal vs. informal credit markets in developing countries. *Journal of Development Economics* 59: 419–444.

Kletzer, K., and B. Wright. 2000. Sovereign debt as intertemporal barter. *American Economic Review* 90: 621–639.

Kochar, A. 1997. An empirical investigation of rationing constraints in rural credit markets in India. *Journal of Development Economics* 53: 339–371.

La Ferrara, E. 2003. Kin groups and reciprocity: A model of credit transactions in Ghana. *American Economic Review* 93: 1730–1750.

Mansuri, G. 2007. Credit Layering in informal financial markets. *Journal of Development Economics* 84: 715–730.

Siamwalla, A., C. Pithong, N. Poapongsakorn, P. Satsanguan, P. Nettayarak, W. Mingmaneenakin, and Y. Tubpun. 1990. The Thai Rural credit system: Public subsidies, private information, and segmented markets. *World Bank Economic Review* 4: 271–296.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71: 383–410.

Udry, C. 1994. Risk and insurance in a rural credit market: An empirical investigation in Northern Nigeria. *Review of Economic Studies* 61: 495–526.

Monocentric Models in Urban Economics

Bruce W. Hamilton

Monocentric models presuppose the existence of a point in geographic space to which access is scarce, hence valuable. They are not concerned with the reasons why access is desirable, but rather with its consequences, especially the manner in which markets allocate access.

Modern monocentric models were developed in the 1960s, largely by Alonso (1964), Muth (1969) and Mills (1972). They are the descendants of the spatial work of von Thünen and the theory of land value due to Ricardo. The land market (perhaps site market is a better term) is the arena in which conflicting demands for access are arbitrated. Site-specific differences in access to the exogeneously located centre are the analogue to Ricardo's differing classes of fertility.

Ricardian theory is agnostic as to the specific reason why some sites are more productive than others, concerning itself with the economic consequences of variations in productivity. It is the work of von Thünen which notes that variation in access is a source of variation in the productivity of land.

The Basic Model

Consider the most familiar monocentric model, in which all production takes place at the central business district (for which CBD is the standard abbreviation), and workers reside in housing that surrounds the CBD. Housing is produced according to the production function

$$x_1 = f(L, K) \tag{1}$$

where x_1 is housing, L is land, and K is capital. Housing requires land, so not all workers can live adjacent to their job; some must commute. As commuting is costly, those who live far from the CBD are worse off than those who live nearby. But if (as we assume) workers have identical preferences and productivity, this utility difference cannot persist in equilibrium, and locationinduced utility differences will be eliminated by the land market.

In the simplest case it is easy to see how this works. Suppose housing demand is perfectly price inelastic, production is by fixed coefficients, and commuting costs \$$t$ per mile. As commuting costs rise the price of a house must fall at \$$t$ per mile to leave all workers with the same utility, and eliminate the excess demand for high-access sites. The price (per building lot) of land must fall at the same rate.

The results are more interesting (and more realistic) when neither production nor demand are fixed coefficients, so we turn to the world of substitution. Our consumer has the utility function

$$V = V[x_1(u), x_2] \tag{2}$$

(where x_1 is housing, as before, and x_2 is all other consumption) and a budget constraint

$$Y - tu = p_1(u)x_1(u) + p_2x_2 \tag{3}$$

where u is distance from the house to the CBD. p_1 has a (u) postscript because the price of housing varies with distance; x_1 has the (u) postscript because a house one mile from the CBD is not the same commodity as a house two miles away. Income is spent on the two goods which generate utility, and on commuting.

Given (2) and (3), we specify a bid–rent function

$$p_1 = p_1(V, p_2, Y, u, t), \tag{4}$$

which is the indirect utility function solved for $p_1(u)$. For given values of V, p_2, Y and t, (4) gives a bid–rent curve, an indifference curve mapped into $p_1 - u$ space. It has the following important property: *If the actual price–distance function is identical to a bid–rent curve, all workers are indifferent among locations.* The following general formula describes the bid–rent curve:

$$\frac{dp_1}{du} = -\frac{t}{x_1} \tag{5}$$

Equation (5) characterizes location equilibrium; if we cross-multiply by x_1 the left-hand side becomes the marginal benefit of decentralization (giving up a unit of access) and the right-hand side is the marginal cost. This function gives exponential decay when p_1x_1 is a constant; that is, when housing demand is unit elastic.

The unit-cost function associated with the production function (1) has as its arguments $R(u)$, the price of land, and i, the cost of capital. By substituting the unit cost function into (4) (replacing $p_1(u)$ with unit cost of housing), we derive a bid–rent function in which the arguments are the price of land (R) and distance, rather than the price of housing and distance.

If we assume only one type of worker, one of the bid–rent curves becomes the equilibrium *housing price–distance function* for the city. We do not know which bid–rent curve represents the equilibrium, because we do not know what utility level the workers can achieve. This will be addressed when we consider the relationship between the city and rest of the world.

Even before completing the model, several interesting properties emerge: (1) The housing price–distance function is concave to the origin, and strictly so if the demand function for housing is not perfectly inelastic. At each distance, the curvature rises monotonically with price elasticity. (2) The land price–distance function is also concave to the origin. At any distance, curvature rises with the price elasticity of demand for housing and with the elasticity of substitution in production. The curvature declines with land's share in production. (3) Both price–distance functions are steeper at any distance, the higher is transport cost. (4) The capital–land ratio falls with distance; since the relative price of land declines with distance, entrepreneurs substitute from capital toward land.

Model Refinements

The Open City

Our city so far has but one type of agent – identical workers who desire access to the CBD so as to economize on commuting. The equilibrium price–distance function must be a bid–rent curve to ensure that workers are indifferent among locations. But each agent has a whole family of bid–rent curves, representing different utility levels. By opening our city to migration we determine which bid–rent curve the workers achieve. Assume this is a small city in a big world and that there is some utility level which workers can achieve elsewhere. Migration occurs so long as workers obtain higher utility here than elsewhere. This drives up housing prices until it is no longer possible to earn a utility premium in this city, and simultaneously determines (1) which bid–rent curve emerges as the rent–distance function, and (2) the size of the city.

We now have a complete model, determining land and housing prices, population density and the capital/land ratio as functions of location; and the size of the city. In addition, this model determines the wage rate; in order to attract a large population, the marginal worker must be compensated for a long commute, and the non-marginal workers for expensive housing. This explains why money wages are higher in big cities than in small ones.

Multiple Sectors

In the simple form we have discussed so far, all agents are identical. To be more realistic we must account for individual variations in both preferences and opportunities. This is the same as allowing different agents to approach the city with different bid–rent curves. A *sector* is defined to include all agents with a given bid–rent curve. Sectors differ from one another according to their bid–rent functions and the alternatives available; that is, by their bid–rent curves. Even if a city contains several sectors, equilibrium requires that all agents from a given sector be on that sector's bid–rent curve. Thus the city's actual rent–distance function is the outer envelope of these sector-specific bid–rent curves; this is the only rent–distance curve which gives no agent an incentive to move – either within the city or to another city.

Several results emerge from this model: (1) The outer-envelope rent–distance curve is concave to the origin (this simply follows from the previously derived result that the individual bid–rent curves are concave to the origin). (2) Sectors are generally geographically segregated – each sector maximizes utility (or profit, as the case may be) only by locating on that portion of the rent–distance function which is its own bid–rent curve. The pattern of segregation has predictable characteristics: A sector is relatively centrally located as (i) it faces high transport cost; (ii) it is non-intensive in land; (iii) its elasticity of substitution between land and capital is large.

Non-CBD Firms

It is widely thought that monocentric models require that all production take place at the CBD (see for example Wheaton 1979). But the assumption is never critical, and it can be relaxed with additions to insight and no loss of tractability. Like the household, the firm has a bid–rent function, whose level curves represent loci of equal profit rather than utility. The level curve representing zero profit is the only one consistent with competition; hence, this is the bid–rent curve which competition forces the firm to bring to the city.

Mills (1972) has an interesting interpretation of the interaction between firm and household location. Suppose the firm produces for export through the CBD. One could imagine workers commuting so production can occur at the CBD. The firm's output does not have to be shipped to the CBD; it is already there. As an alternative, production might be scattered so that jobs are adjacent to workers' homes. Output must be transported to the CBD. But in exchange, workers need not commute. The efficiency question is this: Is it cheaper to transport workers or their output? Mills shows that the market's assignment of sectors to locations works efficiently; firms are assigned to CBD locations if and only if it is cheaper to ship workers than their output. He also notes that over the last century the cost of shipping goods has fallen relatively more than the cost of commuting,

leading to the empirically correct prediction that employment has been decentralizing relative to residential location.

Dynamics

The model described so far is static. The results which describe equilibria assume there are no adjustment costs, or alternatively that the city has been built from scratch during the reign of the current-period parameter values. But in fact, adjustment to a new technology or cost structure entails massive alterations in the capital stock. Structures are among the most durable components of our capital stock, depreciating somewhere between zero and 1 per cent per year (Chinloy 1979). Moreover, depreciation is only part of the story; retirement of structures entails costly demolition or renovation. Mills and Hamilton (1984) find that demolition cost frequently exceeds the value of cleared land, indicating that even fully depreciated structures will not be replaced, but rather abandoned.

Recognition of the empirical importance of adjustment costs has led to a series of attempts to develop dynamic versions of the monocentric models. One approach has been to develop explicitly analytical models; the other to make more or less ad hoc empirical adjustments to the static models. Neither approach has yielded the elegant set of results which emerge from the static model. The analytical models (see Wheaton 1979) quickly become intractable and require detailed assumptions about expectations. The empirical models (see Harrison and Kain 1974) do not have enough generality that we can really see what drives them.

Empirical Investigations

At least in static form, the monocentric models yield a number of definite predictions regarding urban form. (These predictions, along with their derivations and empirical support, are described in Mills and Hamilton 1984.)

In broad outline, these predictions are as follows:

The price of housing declines at a moderate pace with distance (essentially enough to compensate for rising commuting costs). With reasonable parameter values for developed countries, we would expect a decline of about 4 per cent per mile.

The price of land falls rapidly, as the housing price variation described above is driven solely by variation in the price of land. If land's share in housing is 20 per cent (and the elasticity of substitution is unity) we expect land prices to decline about 20 per cent per mile. This of course gives huge spatial variation in land prices within moderate to large cities.

The capital/land ratio falls with distance as the relative price of land falls and entrepreneurs substitute towards land. With a moderate elasticity of substitution, the capital/land ratio should fall about 15 per cent per mile.

Population density falls about 20 per cent per mile. Movement away from the CBD causes households to consume more housing (as it becomes cheaper), and more importantly causes them to consume more land-intensive housing.

The quantitative statements made above are predictions which emerge from the model when we assume reasonable values for parameters. These predictions have been subjected to several empirical investigations.

The first careful empirical study of urban form was carried out by Colin Clark (1951). His findings of regularity in urban population density patterns pre-date the models described in this entry, and must be thought of more as a stimulus to the modellers rather than a test of predictions.

Among formal tests of the models, the first, and still among the most careful, are those carried out by Muth (1969). The results are broadly consistent with the models. Population-density gradients are indeed in the range of 20 per cent per mile; furthermore, these gradients tend to be steeper in nations with high transport costs, and were much steeper in the United States in earlier times. (Mills 1972, has estimated time-trends of population density gradients for the United States, some going back to the 19th century.) In addition, older cities have steeper gradients than newer ones.

The evidence (Chicoine 1981) also supports the prediction of a steep land-price gradient (one consequence of which is that big cities have much higher CBD land values than smaller cities – another prediction which receives empirical support).

Geographic variation in housing prices is predicted to be much smaller than variations in land value or population density; to this extent, empirical studies support the prediction. However, researchers do not uniformly find gradients in the predicted neighbourhood of 4 per cent. Many studies find housing prices rising with distance; other more sophisticated studies find highly irregular housing-price patterns. At this stage we do not know whether these are problems of measurement or genuine failures of the model; see Jackson (1979).

Travel Patterns

A major economic justification for cities is their ability to facilitate inexpensive and rapid interaction among agents. Monocentric models concentrate on one such interaction – the shipment of either workers or goods to the CBD. Yet recent research (Hamilton 1982) has shown that monocentric models do a terrible job of predicting commuting patterns. In a sample of US and Japanese cities the average commute is about eight times that predicted by the models, and is almost as long as is predicted by a model which assumes that the selection of home and job sities is random.

The failure of commuting patterns to conform to the monocentric models should not be surprising. The basic insight underlying monocentric models is that access is valuable, and that the land market serves as the market for this valuable access. But the forms of the monocentric models which have been fully specified have concerned themselves only with access to jobs. In most developed countries, commuting comprises only about 25 per cent of urban travel (unfortunately, little is known about destinations or purposes of other urban trips). The location–equilibrium condition (5) is an excellent concept for thinking about the trade-off between access and the price of housing, but before using this condition to build a specific model of urban form we need to know what access is important. Is it to work, to schools, to friends, to open air, or what? Without knowing the object of one's desire for access, it is impossible to know how to use the location equilibrium condition.

Clearly, access to something is important; otherwise there would be no geographic variation in land values. Perhaps the next generation of urban models will deal more effectively with the richly varied travel demands of modern urban dwellers.

See Also

- ▶ Location of Economic Activity
- ▶ Urban Economics

Bibliography

Alonso, W. 1964. *Location and land use*. Cambridge, MA: Harvard University Press.

Chicoine, D. 1981. Farmland values and the urban fringe: An analysis of sale prices. *Land Economics* 57(3): 353–362.

Chinloy, P. 1979. The estimation of net depreciation rates on housing. *Journal of Urban Economics* 6(4): 432–443.

Clark, C. 1951. Urban population densities. *Journal of the Royal Statistical Society, Series A* 114(4): 490–496.

Hamilton, B. 1982. Wasteful commuting. *Journal of Political Economy* 90(5): 1035–1053.

Harrison, D., and J. Kain. 1974. Cumulative urban growth and urban density functions. *Journal of Urban Economics* 1(1): 61–98.

Jackson, J. 1979. Intraurban variation in the price of housing. *Journal of Urban Economics* 6(4): 464–479.

Mills, E. 1972. *Studies in the structure of the urban economy*. Baltimore: Johns Hopkins Press.

Mills, E., and B. Hamilton. 1984. *Urban economics*. Glencoe: Scott-Foresman.

Muth, R. 1969. *Cities and housing*. Chicago: University of Chicago Press.

Ricardo, D. 1817. *On the principles of political economy and taxation*. London: Bell and sons.

von Thünen, J. 1826. *Der isolierte Staat in Bezheihung auf Landwirtshaft und Nationalekonomie*. New ed. Hamburg, 1863.

Wheaton, W. 1979. Monocentric models of urban land use: Contributions and criticisms. In *Current issues in urban economics*, ed. P. Mieszkowski and M. Straszheim. Baltimore: Johns Hopkins University Press.

Monocentric Versus Polycentric Models in Urban Economics

Tomoya Mori

Abstract

This article overviews the development of the formal modelling framework for the urban spatial structure which started in 1960s and grew dramatically thereafter. Modelling in the 1970s focused on the endogenous formation of the central business district within a city. Then richer polycentric city models were developed in 1980s, where the number, location and spatial extent of the business districts are determined endogenously. The emergence of the new economic geography in 1990s provided a framework capable of explaining the spatial distribution of cities (rather than the business districts within a city) and their industrial structure in a general location- equilibrium model.

Keywords

Urban economics, monocentric versus polycentric models in; Spatial impossibility theorem; Urban agglomeration; Monopolistic competition; New economic geography; Central business district

JEL Classifications

R12; R13; R14; R23; R30

The formal modelling of urban spatial structure originated in the monocentric city model by Alonso (1964). The model was extended to include production, transportation and housing by Mills (1967, 1972) and Muth (1969), and was eventually integrated into a unified framework by Fujita (1989). In these traditional models, the city is a priori assumed to be monocentric, that is, all production activities within a city are supposed to take place in a point representing the *central business district* (CBD), and all workers living in the surrounding area are supposed to commute to the CBD. The success of this model is primarily due to its compatibility with the competitive paradigm, since the existence of the CBD is a priori assumed. In order to explain the urban morphology, however, it is essential to endogenize the CBD formation. For this purpose, Fujita (1986) provided a very useful insight based on the spatial impossibility theorem of Starrett (1978): in order to have endogenous formation of economic agglomeration, the model must have at least one of the following three elements: (*a*) heterogeneous space, (*b*) non-market externalities in production and/or consumption, and (c) imperfectly competitive markets.

The approach based on (a) explains the formation of the CBD by *comparative advantage* among locations, while otherwise retaining the competitive paradigm. One of the earliest such attempts was made by Schweizer et al. (1976).

Most models of type (*b*) are based on *externalities from non-market interactions.* The earliest attempt was made by Solow and Vickrey (1971). In the one-dimensional location space, they considered the optimal allocation of urban land between business areas and roads when each unit of business area is assumed to generate a given number of trips to every other unit. But the first model of residential land use of this type is by Beckmann (1976), where the utility of each individual directly depends on the average distance to all other individuals and the amount of her land consumption. This preference leads to a *bell-shaped spatial population distribution as well as land rent curves*, where the CBD is represented by a densely inhabited area around the central location.

While Beckmann, Solow and Vickrey considered only a single type of agents (firms or consumers), Ogawa and Fujita (1980) and Imai (1982) developed two- sector monocentric models of a one-dimensional city. The dispersion force in this case is generated through land and labour markets. That is, the agglomeration of firms increases the commuting distance for their workers on average, which in turn pushes up the wage rate and land rent around the agglomeration, and this higher cost of labour and land

discourages further agglomeration of firms. The most recent contribution along this line is by Lucas and Rossi-Hansberg (2002), who formally demonstrate the existence of an equilibrium and the endogenous formation of the CBD.

In the endogenous monocentric models discussed so far, the optimal distribution of firms requires greater concentration near the centre than does the equilibrium distribution. The reason is the locational externality generated by individuals: while the location of each individual directly affects the travelling cost for others to make contact with this individual, it is not taken into account when each individual makes a location decision.

Building on Ogawa and Fujita (1980), the first model of a *polycentric city* was developed by Fujita and Ogawa (1982). Their key assumption is that the benefit from interactions between two firms is a negative exponential function of the distance between them, unlike the linear dependence in previous models. When commuting costs are relatively high, this assumption leads to the formation of *multiple business districts* and the possibility of *multiple equilibria.*

The first urban economic model based on (c) is by Fujita (1988). His model demonstrated that pure market interactions alone can explain the agglomeration of economic activities with the use of the Chamberlinian monopolistic competition model. The agglomeration force is generated from the interaction among preference for product variety, transport costs, and increasing returns at the level of individual producers. In this model, the city may be monocentric or polycentric. Also it is possible that business and residential districts are mixed. These works were critical for the emergence of *the new economic geography* (NEG) in the 1990s (Krugman 1991a, b; Fujita 1993).

In the application of the NEG to urban economics initiated by Fujita and Krugman (1995), there are two key features. The first is the *general equilibrium modelling of an entire spatial economy* unlike all the models presented so far. The second is its focus on *the spatial distribution of cities,* while abstracting from the intra-city spatial structure. In particular, it is assumed that mobile firms and workers do not occupy land, so that an agglomeration of firms and population, that is, a city, forms at a point on the continuous location space. The second feature dramatically increases the tractability of the model. The agglomeration force in this model is essentially the same as in Fujita (1988), while the dispersion force is generated from the presence of immobile resources through transport costs between cities and non-city locations. The key to this approach is the recognition that the profitability of any given location for a firm can be represented by an *index of market potential.* The market potential at a given location reflects the trade-off among the proximity to consumers, the degree of competition, and the production cost at that location. In particular, the market potential of a given industry sharply decreases when it moves away from a city in which this industry is agglomerated, and then starts increasing again after a certain distance, exhibiting the presence of an *agglomeration shadow.* Differences in the degree of product differentiation and/or transport costs among industries lead to differences in the size of the agglomeration shadow, which in turn result in variations in the (roughly constant) spacing of agglomerations among industries (Fujita and Mori 1997). In the presence of multiple industries, the *interindustry demand externalities* lead to a formation of *hierarchical city systems* (Fujita et al. 1999). This is reminiscent of Christaller (1933): the set of industries found in a smaller city is a subset of those found in a larger city. Furthermore, the relative decrease in transport costs for urban sectors may eventually lead to the formation of a *megalopolis* consisting of large core cities that are connected by an *industrial belt,* that is, *a continuum of small cities* (Mori 1997). NEG remains the only general location-equilibrium framework which can investigate the spatial distribution of cities and their industrial structure in a unified manner.

There is also a large literature of spatial oligopoly (hence, type *c*) aiming to explain the spatial concentration of stores through *statistical economies of scale*. These models assume that consumers have imperfect information regarding the types (and the prices) of commodities sold by stores before they visit them. The greater the

agglomeration of stores, the more likely it is that consumers will find their favourite commodities. The concentration of stores is explained by the market-size effect due to taste uncertainty and/or lower price expectation (see, for example, Konishi 2005).

Finally, in all the models introduced thus far, all agents are assumed to be atomistic. Hence, land and labour markets are perfectly competitive. In contrast, Henderson and Mitra (1996) offer a model of *suburbanization* in which new *edge cities* are formed by *large land-developers* in the suburbs of the old CBD, formalizing Garreau's observation (1991) on the recent development of edge cities within large US metro areas. Given an existing CBD, the developer of a new edge city chooses the location and capacity of its business district strategically to maximize profits. The developer exercises monopsony power in the labour market in the edge city though her control over aggregate employment there. The proximity to the old CBD increases production efficiency through easier communication of firms between the CBD and the edge city, while it also increases residential land rents and wages of workers in the edge city. This model thus incorporates elements (b) and (c).

See Also

- ▶ Location Theory
- ▶ Spatial Economics
- ▶ Urban Agglomeration
- ▶ Urban Economics
- ▶ Urban Growth
- ▶ Urban Production Externalities
- ▶ Urbanization

Bibliography

Alonso, W. 1964. *Location and land use*. Cambridge, MA: Harvard University Press.

Beckmann, M. 1976. Spatial equilibrium in the dispersed city. In *Mathematical land use theory*, ed. Y. Papageorgiou. Lexington: Lexington Books.

Christaller, W. 1933. *Central Places in Southern Germany*. Trans. C. Baskin. London: Prentice Hall, 1966.

Fujita, M. 1986. Urban land use theory. In *Location theory*, ed. R. Arnott. London: Harwood Academic Publishers.

Fujita, M. 1988. A monopolistic competition model of spatial agglomeration: A differentiated product approach. *Regional Science and Urban Economics* 18: 87–124.

Fujita, M. 1989. *Urban economic theory: Land use and city size*. Cambridge: Cambridge University Press.

Fujita, M. 1993. Monopolistic competition and urban systems. *European Economic Review* 37: 308–315.

Fujita, M., and P. Krugman. 1995. When is the economy monocentric?: von Thünen and Chamberlin unified. *Regional Science and Urban Economics* 25: 505–528.

Fujita, M., and T. Mori. 1997. Structural stability and evolution of urban systems. *Regional Science and Urban Economics* 27: 399–442.

Fujita, M., and H. Ogawa. 1982. Multiple equilibria and structural transition of non- monocentric urban configurations. *Regional Science and Urban Economics* 12: 161–196.

Fujita, M., and J.-F. Thisse. 2002. *Economics of agglomeration: Cities, industrial location, and regional growth*. *Cambridge*: *Cambridge University Press*.

Fujita, M., P. Krugman, and T. Mori. 1999. On the evolution of hierarchical urban systems. *European Economic Review* 43: 209–251.

Garreau, J. 1991. *Edge city: Life on the new frontier*. New York: Doubleday.

Henderson, J., and A. Mitra. 1996. The new urban landscape: Developers and edge cities. *Regional Science and Urban Economics* 26: 613–643.

Imai, H. 1982. CBD hypothesis and economies of agglomeration. *Journal of Economic Theory* 28: 275–299.

Konishi, H. 2005. Concentration of competing retail stores. *Journal of Urban Economics* 58: 488–512.

Krugman, P. 1991a. Increasing returns and economic geography. *Journal of Political Economy* 99: 483–499.

Krugman, P. 1991b. *Geography and trade*. Cambridge, MA: MIT Press.

Lucas, R., and E. Rossi-Hansberg. 2002. On the internal structure of cities. *Econometrica* 70: 1445–1476.

Mills, E. 1967. An aggregative model of resource allocation in a metropolitan area. *American Economic Review* 57: 197–210.

Mills, E. 1972. *Studies in the structure of the urban economy*. Baltimore: Johns Hopkins University Press.

Mori, T. 1997. A modeling of megalopolis formation: The maturing of city systems. *Journal of Urban Economics* 42: 133–157.

Muth, R. 1969. *Cities and housing*. Chicago: University of Chicago Press.

Ogawa, H., and M. Fujita. 1980. Equilibrium land use patterns in a non-monocentric city. *Journal of Regional Science* 20: 455–475.

Schweizer, U., P. Varaiya, and J. Hartwick. 1976. General equilibrium and location theory. *Journal of Urban Economics* 3: 285–303.

Solow, R., and W. Vickrey. 1971. Land use in a long narrow city. *Journal of Economic Theory* 3: 1468–1488.
Starrett, D. 1978. Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 9: 418–448.

Monopolistic Competition

G. C. Archibald

Keywords

Advertising; Chamberlin, E. H.; Characteristics; Discriminating monopoly; Efficient allocation; Entry; Excess capacity; Goods approach versus characteristics approach; Hotelling, H.; Imperfect competition; Increasing returns; Labour market discrimination; Marginal productivity theory; Market failure; Monopolistic competition; Monopsony; Non-convexithy; Ologopoly; Optimality; Planning; Price discrimination; Quasi-rent; Robinson, J. V.; Side-payments; Spatial competition; Tangency solution

JEL Classifications

D4

There is at least an oral tradition that the origin of theories of monopolistic competition is Sraffa's (1926). In the case of Joan Robinson (1933) this may well be true. In the case of Edward Chamberlin (1933) it cannot be: the book was developed from a Ph.D. thesis supervised by Allyn A. Young submitted on 1 April 1927. Indeed, Chamberlin (1933, p. 5 n.) refers to Sraffa's paper as appearing 'since the above was written'.

It is, none the less, convenient to take Sraffa's implicit criticism of Marshall (1890) as a starting point. The increasing-marginal-cost condition, necessary for a competitive equilibrium, was, he asserted, not satisfied in many firms that could not possibly be described as 'Marshallian monopolies'. Thus there existed no appropriate model for an apparently common class of firms (or markets – Sraffa was quite aware of the problems of product heterogeneity). The works of Chamberlin and Mrs Robinson, however diversely prompted, may be seen as attempts to fill what became known as the gap between Marshall's polar cases of monopoly and perfect competition. The gap they had in mind was not filled by oligopoly models, which were already well known. Chamberlin certainly had a 'more competitive' model in mind (free entry). Mrs Robinson was so vague about the construction of the demand curve that it is hard to be sure where 'imperfect competition' leaves off and oligopoly begins, but I read her as in the same spirit as Chamberlin. Whether we can in fact reasonably construct a model of imperfect or monopolistic competition which is not an oligopoly model is still an open question.

The Work of Edward Chamberlin and Joan Robinson

It would not, I think, be a wise use of space to review here the old dispute between Chamberlin (persistent and vociferous) and Mrs Robinson (reluctant and *dégagée*) about whether or not their models were 'the same'. Nor do I wish to dismiss the question as merely 'braces versus suspenders'. Instead, I shall note briefly what it seems to me they had in common and what not. I start with what they had in common.

Both had downward-sloping demand curves (although their construction differed somewhat; see below), but tried to distinguish their models from that of simple or Marshallian monopoly.

This they were able to do because they assumed that the competitive mechanism worked not only through prices but, most importantly, through entry of firms (products). Thus both made an important generalization and extension of Marshall's proposition that competition would ensure that pure profits were only quasi-rents. Indeed, both thought that free entry is a sufficient condition for the elimination of all pure profit in full equilibrium, and thus both exhibited the famous tangency solution.

Thanks to the downward-sloping demand curve, both were able to exhibit profit-maximizing equilibria consistent with non-convexities in the technology, that is to answer Sraffa. (One consequent result, the familiar excess-capacity theorem, is discussed below.)

Both should, in my judgement, be credited with a major extension of the marginal productivity theory of distribution.

There are, none the less, some differences, and they may explain why, in spite of the many elegant features of Mrs Robinson's analysis, Chamberlin's 'monopolistic competition' seems to have been the more enduring model (or, at least, title).

First, there are the famous Chamberlinian 'groups' or industries, groups of similar but not identical products, ill-defined as they may have been. The lack of identity justified the downward slope of the individual demand curve; the assumptions of large numbers and symmetry were carefully stated to justify the assumption of Cournot–Nash behaviour instead of the recognition of oligopolistic interdependence. (The famous construction of the 'perceived' demand curve, DD' and the 'share-of-the-market' demand curve, dd' was designed to explain disequilibrium adjustment behaviour. It has little to do with the properties of full equilibrium which, as in Mrs Robinson's version, is characterized by the elimination of super-normal profit.)

By contrast, Mrs Robinson's treatment of the demand curve seems cavalier. She simply asserted (1933, p. 21) that it shows what the firm will sell at each price when all other adjustments are completed. Whether she had in mind the Cournot–Nash assumption of Chamberlin, or intended to encompass in her model some types of oligopolistic behaviour, is obscure. No adjustment mechanism was suggested. The existence of a full-adjustment demand curve, on which the firm's profit-maximizing decisions are based, was simply postulated as a primitive of the model.

Chamberlin was much more ambitious than Mrs Robinson: he attempted to include product-choice and advertising in the model. I say 'attempted' because it was here that his technique let him down most seriously. Two-dimensional geometry only allowed him to illustrate equilibrium conditions pairwise, and he was never able to exhibit the full set of simultaneous equilibrium conditions, to consider second-order conditions, or to carry out any comparative static analysis. Mrs Robinson confined her attention to what her two-dimensional geometry could handle, omitting advertising and quality from the model, and gave us her elegant analysis of discriminating monopoly and monopsony (with its arresting application to the theory of labour market discrimination).

Criticisms

It would be impossible to review the whole debate over monopolistic competition in limited space. I shall concentrate on those criticisms which seem to be still with us, and lead us to recent advances.

There is no doubt that 'groups' were ill-defined. A common definition, still employed, is that we have a group if we can isolate a set of products such that (i) cross-elasticities of demand between them are 'large' and (ii) cross-elasticities of demand between all members of the set and its complement are 'small'. Triffin (1940) pointed out that there is no analytical cut-off between small and large, and concluded that there was no valid analytical construct between the individual firm and the whole economy. We may take this a little further. We may say that a satisfactory taxonomy induces the discrete metric. A continuous function, such as a cross-elasticity, cannot induce the discrete metric and, accordingly, cannot generate a satisfactory taxonomy. I shall argue below that there now exists an analytically satisfactory way of defining groups, that is, one that induces the discrete metric.

Kaldor (1934, 1935) suggested very early in the discussion that chains of overlapping oligopolies might be empirically more likely than competitive groups operating in virtual isolation from other groups. This raises sharply a question which is still with us: what are the necessary and sufficient conditions for competition to be general, or 'diffuse', that is for the assumption of Cournot–Nash behaviour to be plausible, as opposed to localized or oligopolistic so that the possibility of strategic behaviour has to be admitted.

Several writers on spatial competition have shown recently that free entry cannot be relied upon as a sufficient condition to eliminate supernormal profit, that is to generate the tangency solution (see, for example, Eaton 1976; Eaton and Lipsey 1978). This follows basically from the idea that capital is product- (location-) specific, and long-lasting, and has accordingly to be *committed.* Hotelling (1929) and Chamberlin (1957) thought that monopolistic and spatial competition were, in some sense, the 'same' subject. Given the spatial results, the 'sameness' of the subjects, or models, becomes an urgent question.

Application of Samuelson's (1947) programme, the 'qualitative calculus',to Chamberlin's model, even when 'making the best of it' (to make the criticism more effective) by, for example ignoring the fact that groups were ill-defined, unfortunately showed it to be qualitatively almost empty in the sense of generating few qualitative comparative-static predictions (Archibald 1961). For the individual firm, the reason is the now familiar one: in the multivariate case, the assumption that sufficient extremum conditions are satisfied is not enough to sign the cofactors of off-diagonal elements in the matrix of second-order coefficients. For the group, the reason is essentially the non-convexity of the technology. If, for example demand falls (due, say, to an excise tax), firms exit. When full equilibrium is restored, surviving firms may be producing more or less, that is incurring lower or higher average costs. It also turned out that even the excess capacity theorem did not survive the explicit introduction of advertising in the model (excess capacity remains a possibility but not an entailment). Demsetz (1964) made the interesting suggestion that, by the processes of spin-off, merger, and subcontracting, firms would become so structured that the quantity that minimized average production costs would also minimize average selling costs, in which case equilibrium could not entail excess capacity. It unfortunately turned out that, analytically, this model was inadequately specified (Archibald 1967), but the idea might still be worth pursuing.

Some reactions to the welfare implications of Chamberlin's model were strange. The reaction of several writers to the excess capacity theorem seems to have been 'It can't be true, but, if it is, it is wicked'. Chamberlin replied (1957), quite reasonably, that optimality conditions for an economy with homogeneous product-groups would not necessarily serve as benchmarks for an economy with some increasingness in returns and differentiated product-groups. Little was in fact known about the welfare economics of an economy with non-convexities in its technology.

Some Recent Advances and Unsolved Problems

After some years in which the theory of monopolistic competition was relatively neglected, or at least not much advanced, there has been a recent revival of interest, and a new approach to the subject has emerged. The standard approach, which I shall call the 'goods approach', is in the traditional Walrasian (or Hicksian) style: see the papers by Dixit and Stiglitz (1977), Hart (1979) and Spence (1976). The new, or 'characteristics approach', follows the work of Lancaster: see his (1966, 1971, 1975 and 1979), also Gorman (1980). This approach to monopolistic competition was advocated in Archibald et al. (1986). I note briefly the main features of these two quite distinct approaches.

The goods approach is familiar and traditional, but some features deserve emphasis in the present context. Goods themselves are, of course, the primitives of analysis. There is a fixed vector of possible goods, usually either finite or countably infinite. The utility function is defined on the goods, and there is usually a 'representative consumer' (in some sense that requires definition) who consumes some of each of the goods actually produced. If groups are to be identified, the cross-elasticity taxonomy is employed. If individual firm behaviour is considered, the Cournot–Nash assumption is commonly employed. Full equilibrium is characterized by normal profit.

There are some points to notice here. In some models, the assumption of a fixed vector of goods implies that the technology is not continuous: a firm may choose to produce a good (or quality) x_0 or x_1, say, but cannot produce a good arbitrarily

close to either of them (in some space of attributes). Now, if these attributes (characteristics) of goods are continuous (for example, the fuel consumption of automobiles, the alcohol content of beer), this is a restrictive, and somewhat strange, assumption. Furthermore, it induces an immediate, and perhaps unwelcome, answer to the question, 'are models of monopolistic and spatial competition in some sense the same?', as Hotelling and Chamberlin thought. The space in most spatial models is a continuum, whether in one dimension or two, whence any analogy between the models breaks down at the first step in their construction.

The assumption of a representative consumer who, necessarily, consumes some of each good produced prevents us from taking into account that diversity of tastes which is an obvious feature of the real world. In a characteristics model, the consumer buys no more goods than there are characteristics that he wishes to consume, and if the number of goods produced exceeds the number of 'relevant' characteristics, he buys none of many (perhaps most) goods. This seems to capture an important feature of reality; but it must immediately be admitted that tractable methods of modelling the diversity of preferences have yet to be developed.

The characteristics model is doubtless now familiar too, and only a few points need to be made. The characteristics of goods, rather than the goods themselves, are the primitives of analysis. The technology is assumed to be continuous, in the sense that, if y_1 and y_2 are two goods embodying different mixes of two characteristics, z_1 and z_2 say, then it is possible to produce any good y_i embodying a convex combination of the quantities of z_1 and z_2 embodied in y_1 and y_2. It is thus always possible to produce a good $\in$-close to any other good in the characteristics space. As in spatial models, possible locations form a continuum. Some increasingness of returns is necessarily assumed: with everywhere constant returns, we might expect a 'production point' at every 'consumption point', whether in physical or characteristics space. Thus out of the continuum of possibilities, only a finite number of goods is produced at any time. None the less, it is assumed that, at least in developed economies, the number of goods produced exceeds the number of characteristics desired by consumers. This can only be the consequence of diversity of tastes: if all consumers wanted the same characteristics mix, the number of goods produced would be *less* than the number of characteristics.

An immediate advantage of the characteristics approach is that it allows us to give an analytic definition of a 'group' or industry. It is assumed that the consumption technology is linear, that is, characteristics are 'produced' by goods according to $z = Ay$ where z is the $1 \times m$ vector of characteristics, A is $m \times n$, and y is the $n \times 1$ vector of produced goods. Suppose now that we can partition z, and correspondingly y, so that the corresponding arrangement of the elements of A is block diagonal. Consider one such block, and the corresponding subsets of z and y. We may call this a group: the elements of the subset of y produce only elements of the corresponding subset of z, and no elements of the complement of this subset in y produce any elements of this subset in z.

This taxonomy induces the discrete metric: two goods either do or do not unambiguously belong to the same group. Whether or not there exist, empirically, any groups corresponding to this definition is yet to be discovered.

To complete this sketch of the characteristics approach, let us consider a group of possible goods embodying only two characteristics, z_1 and z_2, say. Then any produced good, y_i say, can be described by θ_i where $\tan \theta_i = z_2/z_1$. The good is completely described by the pair of numbers (p_i, θ_i) where p_i is the price (reciprocal of the length of the vector θ_i to the z_1, z_2 point that can be bought for some fixed amount). The firm's problem is to choose θ_i as well as p_i. The economist's first problem is to characterize the equilibrium vector of θ's as well as p's. His second problem is to characterize the optimal vector of θ's as well as p's. For the first problem, he needs, of course, to know whether competition is oligopolistic or diffuse. (The prior problem of existence has, of course, been thoroughly investigated for the competitive general equilibrium model, and for some partial equilibrium spatial and

small-group models. Little seems to have been done on existence in a Chamberlinian model, at least in the characteristics approach.)

The problem of the socially optimal product choice is of great practical as well as theoretical interest. There is evidence, mostly anecdotal, that the planned economies frequently produce the 'wrong' goods. In the planning literature a fixed vector of homogeneous goods is commonly assumed, and the problem of product choice is not addressed. We similarly lack welfare criteria to tell us if a capitalist economy makes a good job of product selection.

The problem is in fact most difficult. Lancaster (1979) showed that, given some increasingness in returns, considerations of efficiency cannot be successfully divorced from distributional considerations. The problem appears even more starkly in a series of papers by Brown and Heal (1979, 1980, 1981), since they stay with the conventional goods approach. Consider an 'economy' as described by an endowment of resources, a given but non-convex technology, and tastes. They show that, for an arbitrary distribution of ownership, there may exist no efficient allocation of resources. They are also able to show that, for the given 'economy', there always exists a share-ownership distribution (in particular, the equal distribution) such that an efficient allocation does exist.

If this is true for an economy with a fixed product vector, we might conjecture that it is true for an economy in which the product vector is yet to be chosen. We can at least see why efficiency and distribution are entangled in a spatial model (whether the space be geographical or of characteristics). Let there be a given distribution of consumers, whether by location or preferred characteristics mix, and a distribution of stores or products. Assume that the capital specific to one product or location (store) wears out, and is due for replacement. Assume further that some of the mass of demand has shifted (arbitrarily, to the left in the appropriate space). 'Common sense' suggests that the new capital be installed to the left of the old. But this is not a Pareto-efficient move: those consumers remaining to the right will unambiguously lose.

Spence (1976) investigates optimality in monopolistic competition. He adopts the conventional goods model, assuming away income effects, and takes the sum of consumer and producer surplus as his welfare criterion. He is able to show (i) that if sellers can price-discriminate, their profit function will coincide with the welfare maximand, and the optimal product vector (from the possible set) will be produced; and (ii) if not, not: there may be too many or too few products marketed. The reason, roughly speaking, is that, with some increasingness of returns, and products to be chosen, price is not a sufficient signal: we have a species of market failure. Thus there is no market in which you and I and the producers may arrange side-payments such that, by agreeing on the same good(s), we all benefit from the increasingness of returns.

I conjecture that Spence has given us all the pure efficiency results that are to be had. If we follow Lancaster, and Brown and Heal, and do not ignore distributional considerations, it is not obvious what results we may hope for.

We urgently need to know the necessary and sufficient conditions for competition to be diffuse (Chamberlinian) as opposed to local or oligopolistic. The next question follows: on what conditions does the small-group model become asymptotically competitive? So far, we have only a scattering of results.

Consider the set of products in a space of two characteristics, or of stores along a line. It is obvious that no product, or store, can have more than two neighbours: we appear to have Kaldor's chain of overlapping oligopolies. What happens if we increase the number of consumers and products, or stores, without limit is less obvious. While each outlet still has no more than two neighbours, its scope for price setting is evidently diminished. We might conjecture that the asymptotic results in this case will be approximately competitive. Now let the dimensions of the tangency solution space increase. It has been shown, by Archibald and Rosenbluth (1975), that when the number of tangency solution is four, the number of neighbours (immediate competitors) each product may have approaches half the number of products in the

space. This is a necessary condition for competition among diverse products to be Chamberlinian. Sufficient conditions have not been established. These authors, and others, also considered the possibility of 'pre-emptive entry': an incumbent firm in a growing market occupies a point (in physical or characteristics space) before it is normally profitable to do so in order to deter new competition.

Hart (1979) gets asymptotically competitive results in a goods model. He assumes, however that the output of each firm is bounded from above so that the output of each firm can be made as small as we like relative to the whole economy. Further, replication involves increasing the number of consumers each of whom has one of a finite set of preferences, that is, cloning them. What is not yet known is what happens asymptotically in an economy in which (i) the output of the individual firm is not bounded, (ii) the 'address' of products, in the sense of Archibald, Eaton and Lipsey matters, and (iii) as the number of consumers increases, so does the diversity of preferences.

See Also

- ▶ Advertising
- ▶ Chamberlin, Edward Hastings (1899–1967)
- ▶ Competition
- ▶ Market Structure
- ▶ Oligopoly
- ▶ Product Differentiation
- ▶ Robinson, Joan Violet (1903–1983)

Bibliography

Archibald, G.C. 1961. Chamberlin versus Chicago. *Review of Economic Studies* 24: 9–28.

Archibald, G.C. 1967. Monopolistic competition and returns to scale. *Economic Journal* 77: 405–412.

Archibald, G.C., and G. Rosenbluth. 1975. The 'new' theory of consumer demand and monopolistic competition. *Quarterly Journal of Economics* 80: 569–590.

Archibald, G.C., B.C. Eaton, and R.G. Lipsey. 1986. Address models of value theory. In *New developments in the analysis of market structure*, ed. J.E. Stiglitz. Cambridge, MA: MIT Press.

Brown, D.J., and G. Heal. 1979. Equity, efficiency and increasing returns. *Review of Economic Studies* 57: 571–585.

Brown, D.J., and G.M. Heal. 1980. Two-part tariffs, marginal cost pricing and increasing returns in a general equilibrium model. *Journal of Public Economics* 13: 25–49.

Brown, D.J., and G. Heal. 1981. *Welfare theorems for economies with increasing returns*, Essex economic papers No. 179.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA/London: Harvard University Press/Oxford University Press.

Chamberlin, E.H. 1957. *Towards a more general theory of value*. New York: Oxford University Press.

Demsetz, H. 1964. The welfare and empirical implications of monopolistic competition. *Economic Journal* 74: 623–641.

Dixit, A.K., and J.E. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Eaton, B.C. 1976. Free entry in one dimensional models. *Journal of Regional Science* 16: 21–33.

Eaton, B.C., and R.G. Lipsey. 1978. Freedom of entry and the existence of pure profit. *Economic Journal* 88: 455–469.

Gorman, W.M. 1980. A possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47: 843–857.

Hart, O. 1979. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies* 46: 1–30.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39: 41–57.

Kaldor, N. 1934. Mrs. Robinson's 'economics of imperfect competition'. *Economica NS* 1: 335–341.

Kaldor, N. 1935. Market imperfections and excess capacity. *Economica NS* 2: 33–50.

Lancaster, K.J. 1966. A new approach to consumer theory. *Journal of Political Economy* 74: 132–157.

Lancaster, K. 1971. *Consumer demand: A new approach*. New York: Columbia University Press.

Lancaster, K. 1975. Socially optimal product differentiation. *American Economic Review* 65: 567–585.

Lancaster, K.J. 1979. *Variety, equity, and efficiency*. New York: Columbia University Press.

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

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Spence, A.M. 1976. Product selection, fixed costs, and monopolistic competition. *Review of Economic Studies* 43: 217–235.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36: 535–550.

Triffin, R. 1940. *Monopolistic competition and general equilibrium theory*. Cambridge, MA: Harvard University Press.

Monopolistic Competition and General Equilibrium

Takashi Negishi

Traditional general equilibrium theory, as exemplified in Walras (1874–7) and Hicks (1939), was concerned only with perfect competition, though it was preceded by Cournot's theory of oligopoly (1838), where perfect competition is only a limiting case of oligopoly. Walras (1874–7, p. 431) admitted that perfect competition is not the only possible system of economic organization and that we must consider the effects of other systems, such as those of monopolies, in order to make a choice between perfect competition and the other systems, as well as to satisfy our scientific curiosity. His theory of monopoly, however, remains a partial equilibrium analysis and no general equilibrium model is developed for an economy which contains monopolies. Hicks was more explicit in excluding monopolies from general equilibrium theory. He insisted that 'a universal adoption of the assumption of monopoly, must have very destructive consequences for economic theory' (1939, p. 83). The effect of an increase in demand on price is indeterminate, if the expansion of the firm is stopped not by rising costs, as in the case of competition, but by the limitation of the market, as in the case of monopoly.

But it is exactly on this problem of the rising costs versus the limitation of demand that Sraffa (1926) based his arguments against the empirical relevance of perfect competition. He argued that the chief obstacle to increasing production 'does not lie in the cost of production but in the difficulty of selling the larger quantity of goods without reducing the price' (p. 543). Although the theory for firms facing downwardly sloping demand curves suggested by Sraffa was first developed by Chamberlin (1933) and Robinson (1933) within the framework of Marshallian partial equilibrium theory, Triffin (1940, p. 89) emphasized that 'the new wine of monopolistic competition should not be poured into the old goatskins of particular equilibrium methodology'. For Triffin, the main contribution of monopolistic competition theory lay in its focus on the interdependence of firms. Since partial or particular equilibrium theory deals only with relations between firms in an industry (or a group), the general theory of economic interdependence has to be constructed so as to encompass interrelations among all firms in an economy. Modern theories of monopolistic competition and general equilibrium such as Negishi (1961), Kuenne (1967), Arrow and Hahn (1971), Gabszewicz and Vial (1972), Fitzroy (1974), Marschak and Selten (1974) should be seen in this historical perspective.

Let us start with the simple model of monopolistic competition and general equilibrium considered by Negishi (1961). Suppose an economy is composed of perfectly competitive consumers, perfectly competitive firms and monopolistically competitive firms. As usual, a perfectly competitive firm is assumed to maximize its profit by choosing a combination of input (vector) and output (vector) from a technologically given convex set of feasible combinations of input and output when prices are given. Similarly, a perfectly competitive consumer is assumed to maximize his utility subject to budget constraint, when prices and the distribution of profit from all firms are given.

A monopolistically competitive firm is assumed to perceive a *subjective* inverse demand curve when a currently observed combination of price and quantity is given. A subjective demand curve is different from an objective or true demand curve, which is a locus of actually realized or observable combinations of price and quantity. Suppose customers actually demand x^* of its product at the price of p^* Then the firm perceives a possible relation between the demand x and the price p such that

$$p^* = D(x^*, p^*, x^*), \quad \partial p/\partial x < 0. \tag{1}$$

To make the perceived demand curve rational in the weakest sense, the condition that

$$p^* = D(x^*, p^*, x^*) \tag{2}$$

must be imposed, so that the perceived demand curve passes the observed point (p^*, x^*). In other words, it intersects the objective demand curve at the given observed point. If (1) is simplified so that p is linear with respect to x; that is,

$$p = a(p^* - x^*) - b(p^*, x^*)x \tag{3}$$

where a and b are positive, the profit of the firm is a concave quadratic function of inputs and outputs. The firm is assumed to maximize its profit by choosing a combination of outputs and inputs when the currently observed combination of quantity demanded and price is given in markets where the firm is a price-maker, and prices are given in all the other markets.

Finally, all the markets must be cleared. In products markets the quantity demanded by consumers and by all the firms as input has to be equal to the quantity of output of all the supplying firms. In markets of factors of production, the quantity demanded by all the firms as input has to be equated to the quantity supplied by consumers. Price is raised (lowered) if quantity demanded exceeds (falls short of) quantity supplied, which we will call here the law of supply and demand.

Suppose we are given a set of values for all the prices, the inputs and outputs of all the firms, and quantities demanded and supplied for all the consumers. Then through the behaviour of markets, consumers and firms, there will be generated a new set of values of these prices and quantities corresponding to the given original set. Firms choose new combinations of inputs and outputs so as to maximize profits in view of given prices and quantities, while consumers choose new quantities demanded and supplied so as to maximize utilities, since the profits distributed are calculated from given prices and quantities. A new vector of prices is generated in markets through the law of demand and supply. Generally, the new vector of prices and quantities is different from the original. Under standard technical assumptions, however, we can show by the use of Kakutani's fixed-point theorem that there exists a vector completely identical to itself.

It is easily seen that such an unchanging vector of prices and quantities represents a general equilibrium of an economy which contains monopolistic competition. Since prices are unchanging, all the markets are cleared with quantities demanded and supplied, chosen by utility and profit maximization. Since price and quantity are unchanged, we see from (2) that perceived profit of a monopolistically competitive firm is maximized at the observed, realized price and quantity demanded; that is, the firm perceives demand for its product correctly. In other words, the existence of a general equilibrium is proved for an economy with monopolistic competition.

Certainly the model described above has many unsatisfactory aspects, and many criticisms, modifications and generalization have been suggested, some of which are reviewed below.

The increasing returns to scale is presumably one reason for the existence of monopoly and monopolistic competition. Therefore, Arrow and Hahn (1971, pp. 151–67), Fitzroy (1974), Silvestre (1977, 1978) and others have emphasized the importance of the case where the feasible set of combinations of input and output is not convex for monopolistically competitive firms, and have developed interesting models to deal with this problem. For example, the model considered by Arrow and Hahn is very general with respect to the behaviour of monopolistically competitive firms, since each firm's reaction function is simply assumed to be continuous with respect to relevant variables, and the maximization of profit is not explicitly considered.

This is not unrelated to the objection against profit maximization raised by Gabszewicz and Vial (1972) for the case of monopolistically competitive firms. The owners of a firm may be interested not in profit itself but rather in what the profit can buy. The owners of price-making firms may, then, prefer a lower profit but favourable prices for consumption goods, to higher profit and unfavourable prices. To some extent, however, this difficulty has been solved by Hart (1982, 1984).

Nikaido (1975, pp. 7–10) was very critical of the use of the perceived or subjective demand curve (1) on the grounds: (a) that in monopolistically competitive markets disequilibrium does not consist in excess demand or supply (Lange 1944, p. 35), and (b) that monopolistically competitive

firms must perceive demand correctly not only at equilibrium (which is guaranteed by condition (2)) but also at disequilibria. Confining himself to the use of the Leontief model of production and additive logarithmic utility functions, however, Nikaido also found difficulties in the construction of the objective demand curve, which may not be downward sloping (1975, pp. 53–6).

Hart's (1984) criticism on the use of the subjective demand curve (1) and (2) is that the class of possible equilibria is very large so that the model gives us very little predictive power. In this respect, his consideration of the case of the reasonable conjecture (Hahn 1978) is very interesting, since it reduces the class of possible equilibria by imposing restrictions more stringent than (2).

Unlike the case of the subjective or perceived demand curve (1), the objective demand curve is derived explicitly from the behaviour of consumers; that is, utility maximization. Models with monopolistically competitive firms facing such objective demand curves were developed first by Gabszewicz and Vial (1972) and Fitzroy (1974) on the basis of the *Cournot–Nash equilibrium* concept. They were followed by Marschak and Selten (1974), Laffont and Laroque (1976), Silvestre (1978) and others, who contributed to an interesting development of concepts of equilibrium.

In the model with the subjective demand curve, the restriction (3) is imposed in order that the profit function of a monopolistically competitive firm be concave with respect to the level of output, so that the level of output becomes a continuous function of the given values of prices and quantities. Similarly, in models with objective demand curves corresponding conditions must be imposed to make the profit function concave and the reaction function continuous, for otherwise it is difficult to prove the existence of a general equilibrium. However, Roberts and Sonnenschein (1977) have produced a number of non-pathological examples where these conditions are not satisfied, and have argued for the non-desirability of imposing any such conditions that are not derived from hypotheses on the fundamental data of preferences, endowments and technology. Therefore, it cannot be said that the problem of the existence of a general equilibrium is solved satisfactorily for an economy with monopolistically competitive firms facing objective demand curves.

Perhaps Kuenne's criticism (1967) deserves a separate mention. He argued against Triffin's interpretation (1940) of the relation between Walras–Pareto general equilibrium theory and Chamberlin's theory of monopolistic competition, and criticized the model of Negishi (1961) on the ground that it does not cope with the problem of product differentiation, interproduct competition being eliminated in both its rivalrous and non-rivalrous aspects. He developed a general equilibrium model to study interrelated product markets, by adopting the assumption of non-rivalrous inter-firm competition and by employing the concept of industry and group in the sense of Chamberlin.

Recently an interesting model was constructed by Dixit and Stiglitz (1977) to study the problem of product differentiation in the spirit of Chamberlin. Following Krugman (1979), we may sketch a simplified version of their model as follows. Let us consider an economy with only one scarce factor of production, labour. The number of products differentiated is a variable which is denoted by n. The utility function of the representative consumer, into which all products enter symmetrically, is

$$U = \sum v(c_i), \qquad v' > 0, v'' < 0, \tag{4}$$

where c_i is the consumption of the ith product. All products are assumed to be produced with the same cost function. The labour used in producing each product is a linear function of output:

$$y_i = a + bx_i \tag{5}$$

where y_i is labour used in production of the ith product, x_i is the output of the ith product, and a and b are positive constants. Since all the products are symmetric, it follows that $x_i = x$, $y_i = y$, $c_i = c$, for all i. If c is known from the condition that the maximized profit is zero, the number of products differentiated is obtained from (5) as

$$n = L(a + bx) \tag{6}$$

since $x = Lc$ and $ny = L$, where L is the given labour force.

We cannot discuss here in detail all the critical arguments and suggestions for new concepts cited above, nor can our survey be exhaustive of the rapidly growing literature; fortunately some of it is nicely surveyed in Hart (1984).

We have seen that there is little agreement achieved among scholars on how monopolistic competition should be modelled in general equilibrium theory. Many scholars are critical of the model that uses the subjective demand curve perceived with weak consistency conditions and try to develop models using the objective demand curve derived explicitly from the utility maximization of consumers. In other words, they insist that monopolistically competitive firms should be modelled as rational agents fully informed of market conditions summarized in objective demand curves. We do not deny, of course, the importance of the problem of whether the behaviour of such fully informed rational firms is mutually consistent in the sense that there exists an equilibrium in the model using an objective demand curve. Unfortunately, it is difficult to solve this problem in view of Nikaido (1975) and Roberts and Sonnenschein (1977).

In the case of a perfectly competitive economy, however, it is only recently that existence problems were solved for a model in which agents are fully informed of market conditions. But long before that, there were already important and useful applications of general equilibrium theory to many problems, where the existence of an equilibrium was assumed as a part of hypotheses (Hicks 1983, p. 374). Similarly, there have already been several interesting applications of the theory of general equilibrium with monopolistic competition, even though no particular model has won general acceptance. Since there is no reason why different models should not be used for different applications, however, we can use a model with the subjective demand curve or a model with very strong assumptions, provided that it yields an interesting result in a particular field.

The theory of international trade is set apart from other parts of economics by its concern with general equilibrium. Since scale economies play a crucial role in explaining the postwar growth in trade among the industrial countries (Kaldor 1966; Balassa 1967), general equilibrium theory with monopolistic competition should be applied in order to deal with those problems of increasing returns that cannot be dealt with by the theory of the perfect competition. A representative contribution in this area is Krugman (1979), who showed that trade is a way of extending the market and allowing exploitation of scale economies, and need not be a result only of international differences in technology or factor endowments. This is a revival of Adam Smith's argument that the division of labour is limited by the extent of the market (1776, p. 31). Though he used a Dixit–Stiglitz model, Krugman also surveyed related contributions based on other models of monopolistic competition.

Recent literature on Keynesian economics considers Keynesian equilibria by assuming that prices and wages are fixed, and effective demands and supplies equilibrated through the adjustment of quantities. One problem here is why prices and wages are fixed in the face of the existence of involuntary unemployment and excess capacities or inventories. Since the theory of perfect competition cannot solve this problem, it is natural to consider applications of the theory of monopolistic competition. An interesting example of the rapidly growing literature on this topic is Hart (1982), which also contains references to other contributions.

Having assumed that firms know the objective demand curves facing them, Hart (1982) had to admit a serious nonexistence problem pointed out by Roberts and Sonnenschein (1977). However, it is particularly in Keynesian economics that we should use subjective demand curves. Unlike the case of Walrasian homogeneous markets, where agents are fully informed of market conditions, markets in Keynesian economics should be Marshallian heterogeneous markets, where agents are not fully informed of conditions necessary to know the objective demand curves. In such a market, even competitive firms cannot perceive the demand curves to be infinitely elastic if demand falls short of supply. This is the reason why prices are fixed in the face of excess supplies,

and demand and supply are equilibrated through the adjustment of the supply.

There is a strong reason why the perceived demand curve has a kink at the currently realized point, due to asymmetric behaviour of consumers in a world of imperfect information. Competitive firms cannot exceed the current price ruling in the market, since a higher price would induce customers to search for low-price suppliers. The perceived demand curve is infinitely elastic to the left of the currently realized point. A lower price, on the other hand, may not be fully advertised to customers who are currently buying from firms that are not lowering their prices. The perceived demand curve is, therefore, rather inelastic to the right of the realized point. The existence of this kink makes the profit function of the firm concave with respect to the level of the output, which is convenient for proving the existence of an equilibrium (Negishi 1979).

See Also

- ▶ Conjectural Equilibria
- ▶ Cooperative Equilibrium
- ▶ Nash Equilibrium
- ▶ Oligopoly and Game Theory

Bibliography

Arrow, K.J., and F. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Balassa, B. 1967. *Trade liberalization among industrial countries*. New York: McGraw-Hill.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Cournot, A.A. 1838. Recherches sur les principes mathématiques de la théorie des richesses. Trans. N.T. Bacon. New York: Macmillan, 1897.

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67(3): 297–308.

Fitzroy, F. 1974. Monopolistic equilibrium, non-convexity and inverse demand. *Journal of Economic Theory* 7(1): 1–16.

Gabszewicz, J.J., and J. Vial. 1972. Oligopoly 'à la Cournot' in a general equilibrium analysis. *Journal of Economic Theory* 4(3): 381–400.

Hahn, F.H. 1978. On non-Walrasian equilibria. *Review of Economic Studies* 45(1): 1–17.

Hart, O.D. 1982. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics* 97(1): 109–138.

Hart, O.D. 1984. Imperfect competition in general equilibrium: An overview of recent work. In *Frontiers of economics*, ed. K.J. Arrow and S. Honkapohja. Oxford: Blackwell.

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

Hicks, J.R. 1983. *Classics and moderns. Collected essays on economic theory*, vol. III. Oxford: Blackwell.

Kaldor, N. 1966. *Causes of the slow rate of economic growth in the United Kingdom*. Cambridge: Cambridge University Press.

Krugman, P.R. 1979. Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9(4): 469–479.

Kuenne, R.E. 1967. Quality space, interproduct competition, and general equilibrium theory. In *Monopolistic competition theory: Studies in impact*, ed. R.E. Kuenne. New York: J. Wiley.

Laffont, J.J., and G. Laroque. 1976. Existence d'un équilibre général de concurrence imparfaite: une introduction. *Econometrica* 44(2): 283–294.

Lange, O. 1944. *Price flexibility and employment*, Cowles Commission Monograph, vol. 8. Bloomington: Principia Press.

Marschak, T., and R. Selten. 1974. *General equilibrium with price-making firms*, Lecture Notes in Economics and Mathematical Systems, vol. 91. Berlin: Springer.

Negishi, T. 1961. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.

Negishi, T. 1979. *Microeconomic foundations of Keynesian macroeconomics*. Amsterdam: North- Holland.

Nikaido, H. 1975. *Monopolistic competition and effective demand*. Princeton: Princeton University Press.

Roberts, J., and H. Sonnenschein. 1977. On the foundations of the theory of monopolistic competition. *Econometrica* 45(19): 101–113.

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Silvestre, J. 1977. General monopolistic equilibrium under nonconvexities. *International Economic Review* 18(2): 425–434.

Silvestre, J. 1978. Increasing returns in general noncompetitive analysis. *Econometrica* 46(2): 397–402.

Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. R.H. Campbell, A.S. Skinner, and W.B. Todd. Oxford: Oxford University Press, 1976.

Sraffa, P. 1926. The laws of returns under competitive conditions. *Economic Journal* 36: 535–550.

Triffin, R. 1940. *Monopolistic competition and general equilibrium theory*. Cambridge, MA: Harvard University Press.

Walras, L. 1874–7. *Éléments d'économie politique pure ou théorie de la richesse sociale.* Definitive edn, Lausanne: Corbaz. Trans. W. Jaffé. Homewood: Richard D. Irwin, 1954.

Monopoly

Edwin G. West

JEL Classifications
D4

Irving Fisher (1923), once defined monopoly simply as an 'absence of competition'. From this point of view various attitudes to, or criticisms of, monopoly are connected with the particular vision of competition that each writer has in mind. To the neoclassical economist monopoly is the polar opposite to the now familiar 'perfect competition' of the textbooks. Modern writers in the classical tradition, on the other hand, complain that perfect competition neglects the *process* of competitive activity, overlooks the importance of time to competitive processes and assumes away transaction or information costs.

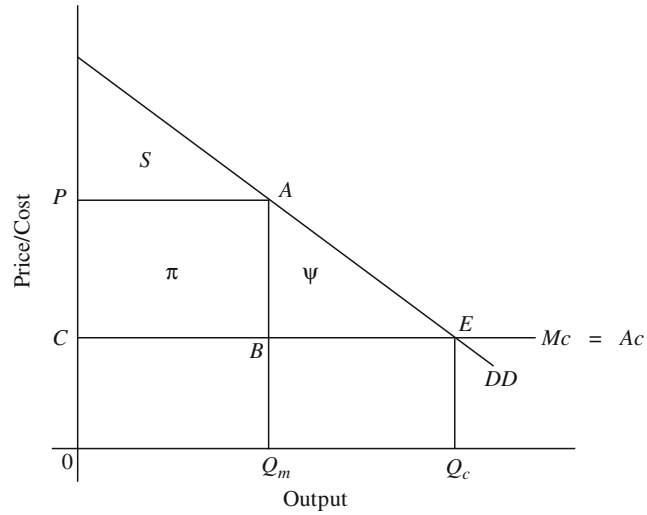
In effect, 'perfect competition' to the neoclassical implies perfect decentralization wherein exchange costs happen to be zero. But the modern critics insist that exchange is not costless. And for this reason competition can be consistent with a wide variety of institutions that are employed to accommodate time, uncertainty and the costs of transacting (Demsetz 1982). Such arrangements include, for example, tie-in sales, vertical integration and manufacturer-sponsored resale price maintenance. Such price-making behaviour means that in the real world decentralization is imperfect. And it is imperfect decentralization that is embodied in the classical paradigm of laissez faire. Consequently many phenomena that are automatically treated by the neoclassical as the absence of perfect competition or the presence of behaviour that *looks* monopolistic, are often viewed approvingly by those in the classical tradition.

It is widely believed that, historically, Adam Smith's *Wealth of Nations* provided the most sustained and devastating attack on monopoly. It is true that he speaks of 'monopoly' quite frequently, but typically he uses the term in a wide 18th-century sense to include all kinds of political restrictions. Monopoly under the modern meaning of a single uncontested firm was not Smith's usual target. He employed the term most often to refer to multi-firm industries enjoying statutory protection. Thus, 'the law gave a monopoly to our boot-makers and shoe-makers, not only against our graziers, but against our tanners' (Smith (1776), 1960, vol. 2, p. 153). Again, the whole system of mercantilism was condemned as monopolistic: 'Monopoly of one kind or another, indeed, seems to be the sole engine of the mercantile system' (ibid., p. 129).

The Ricardians too were more concerned with general restrictions, and especially with the fixed supply of land. Ricardo's *Principles of Political Economy and Taxation* in fact has only five pages out of 292 that discuss monopoly, while John Stuart Mill's *Principles of Political Economy* has only two out of 1,004. Following the Ricardians, the development of Darwinian philosophy in the mid-19th century only served to reinforce the classical emphasis on the necessity, if not inevitability, of competition. It is true that the 'modern' and more rigorous theory of monopoly, showing equilibrium to be determined by the equality of marginal revenue with marginal cost, was introduced by Cournot in 1838. But it received very little attention until much later.

In America the classical laissez-faire view of competition and imperfect decentralization prevailed at least to the end of the 19th century. When the Sherman Antitrust Act was passed in 1890, economists were almost unanimously opposed to it. Thus, despite his general disposition for widespread government intervention, the founder of the American Economic Association, Richard T. Ely (1900), firmly rejected the politically popular policy of 'trust busting'. In the late 1880s John Bates Clark similarly feared that antitrust laws would involve a loss of the efficiency advantages of combinations or trusts. Combination itself was often necessary to generate adequate capital and to insure against adversity during the depressing period of the business cycle. Other contemporary economists, including Simon N. Patten, David A. Wells and George Gunton, had similar views. The last argued that

Monopoly, Fig. 1



the concentration of capital does not drive small producers out of business, 'but simply integrates them into a larger and more complex system of production, in which they are enabled to produce wealth more cheaply for the community and obtain a larger income for themselves'. Instead of the concentration of capital tending to destroy competition, the reverse was true: 'By the use of large capital, improved machinery and better facilities, the trust can and does undersell the corporation' (Gunton 1888, p. 385).

Consider now, and in contrast, the subsequent neoclassical approach which eventually involved the comparison of monopoly with what is said to be its polar opposite market structure of perfect competition. The method was gradually developed from the last part of the 19th century and ultimately, in the 1950s, reached the stage of empirical measurement of what was described as the social cost of monopoly. The most influential study has been that of Harberger (1954), whose basic argument can be summarized in terms of Fig. 1.

Assume that long-run average costs are constant for both firm and industry and are represented by the line $M_c = A_c$. The perfectly competitive output would be at Q_c where M_c intersects the demand curve *DD*. If a monopolist were substituted, he could maximize profits by producing Q_m at price *P*. His monopoly profit, π, would be represented by the rectangle *ABCP*. The loss of consumers' surplus is measured by the trapezoid *AECP*. The part of this area represented by *ABCP*, however, is not destroyed welfare but simply a transfer of wealth from consumers to the monopolist. The net loss to society as a whole from the monopoly is given by the 'welfare triangle' *ABE*, denoted in Fig. 1 by ω. After making some heroic assumptions, in particular that marginal cost (M_c) was constant for all industries and that the price elasticity of demand was unity everywhere, Harberger estimated an annual welfare loss of $59 million for the US manufacturing sector in the 1920s. This figure was surprisingly small since it represented only one-tenth of 1 per cent of the US national income for that period.

Subsequent writers have argued that Harberger's measure was a serious underestimate for statistical and other reasons. George Stigler (1956) objected that (1) monopolists normally produce in the range where elasticity is greater than unity; (2) some monopoly advantages become embodied in the accounted costs of assets, so leading to an underestimate in reported profits. Subsequent studies that allowed for Stigler's objections reported social costs of monopoly much higher than Harberger's. Thus D.R. Kamerschen (1966) reported an annual welfare loss due to monopoly in the 1956–61 period amounting to around 6 per cent of national income. D.A. Worcester, Jr. (1973), on

the other hand, using *firm* rather than *industry* data, and assuming an elasticity of (minus) 2, reported a maximum estimate of welfare loss in the range of 0.5 per cent of national income for the period 1965–9. Focusing on the complaint that Harberger assumed the normal competitive profit rate to be represented by the actual average profit rate earned, whereas the latter itself contains a monopoly profit element, Cowling and Mueller (1978), reported that 734 large firms in the US generated welfare losses totalling $15 billion annually over the period 1963–6, and this amounted to 13 per cent of Gross Corporate Product. All such criticisms have obviously been of a technical nature and implicitly accept Harberger's basic methodology.

Consider next another type of qualification that also accepts the same central methodology. In the frictionless world of the neoclassical model, where all exchange costs are zero, it would be profitable for the monopolist to produce more than Q_m in Fig. 1. This would be the case, for example, with the institution of a two-part tariff where a second price is charged for all purchases in excess of Q_m. If this price were located exactly halfway between P and C, it could be shown that the triangle of welfare loss would shrink to one-quarter of the existing size of ω. An extension of such multi-part pricing, of course, would reduce the welfare triangle of loss still further. With the presence of zero exchange costs, which pertains to the neoclassical world, perfect price discrimination is possible. In this case the whole of the trapezoid *CPAE* would consist of transferred wealth from consumers to producers. Deadweight welfare loss from monopoly would be zero.

If the neoclassical analyst objects that perfect price discrimination does not exist in the real world, he has to offer reasons. It is difficult, meanwhile, to conceive of any practical explanation that could be couched in terms of anything else but significant costs of exchange, such as positive information costs and risk. But such explanation undermines the 'purity' of the neoclassical model and points us back in the direction of the classical world of imperfect decentralization featuring real-world limitations on knowledge, and the existence of dynamic change under uncertainty.

It will be helpful now to describe classical analysis in terms of Fig. 1. But first recall that, instead of the notion of perfect competition as a static long-term equilibrium, we start with the view of competition, espoused by Adam Smith and his successors, as a process of rivalry within a time dimension. In Schumpeter, for instance, competition is seen as 'a perennial gale of creative destruction'. It is the possibility of profit, of course, that drives the innovating entrepreneur. Without it the laissez-faire model of decentralization would collapse. But once profits are obtained by a successful pioneer his operation is immediately copied by others, so that there is a constant tendency for entrepreneurial profit to be competed away. It is this focus on a continual series of short runs that distinguishes the analysis from that of 'perfect competition', which is always expressed in terms of the very long run.

Assume then the discovery of a new product, product X, by an entrepreneur who proceeds to offer Q_m of it at price P (see Fig. 1). It is only academically true that he is restricting output compared with what potential rivals would produce if they possessed his knowledge and business acumen. But since, in reality, they do not, the only alternative to Q_m supply of product X is some positive quantity of conventional products that the factors were previously producing (the supply of X being zero). The result of his activity in producing X, therefore, is pure social gain, and this is measured in Fig. 1 by the profit plus the consumer surplus S. The welfare triangle of social loss (ω) does not exist. It can be expected that the entrepreneur's action will lead to the eventual entry of rivals. At this stage competition will lead to a lowering of price towards cost. This process will then involve a transfer of wealth from the original entrepreneur to consumers. But the latter's original and temporary profit is necessary to induce him to introduce the product at an earlier time than otherwise. It is this earlier introduction indeed that produces the social gains. So while such temporary profit may be described as proceeding from the market structure of 'imperfect competition', nevertheless, according to the Smithian/Schumpeterian analysis, the monopolies so described are necessary institutions, since economic growth would be

much weaker without them. Indeed, society recognizes such logic when it grants temporary legal monopolies in the form of patents.

It is necessary now to examine the special place that is usually accorded to the phenomenon of what is called 'natural monopoly'. This is said to exist when it is technically more efficient to have a single producer or enterprise. The ultimate survival of such a single firm is usually the natural outcome of initial rivalry between several competitors. J.S. Mill (1848, 1965, p. 962) appears to have been the first to use the adjective 'natural' and to use it interchangeably with 'practical'. Examples quoted by Mill included gas supply, water supply, roads, canals and railways.

In his *Social Economics* (1914) Friedrich von Wieser was probably the first to distinguish the modern from the classical doctrine of monopoly. The classical (Marxian?) attribution to monopoly of the 'favoured' market position of capital over labour was incorrect. So was Ricardo's reference to the 'monopoly' of agricultural soil. The price of urban rents was a competitive price. A typical real monopoly for von Wieser consisted of what he called the 'single-unit enterprise', that was identical to the organization that Mill had previously identified as a 'natural monopoly'. The postal service was an excellent illustration:

> In the face of [such] single-unit administration, the principle of competition becomes utterly abortive. The parallel network of another postal organization, beside the one already functioning, would be economically absurd; enormous amounts of money for plant and management would have to be expended for no purpose whatever (von Wieser (1914), 1967, pp. 216–17)

The conclusion was that some kind of government control such as price regulation was required.

One must conjecture that von Wieser would have been astonished by the application (in the 1980s) to natural monopolies of the new theory of 'the contestable market'. According to its promulgators, this is a situation in which 'entry is absolutely free, and exit is absolutely costless' (Baumol 1982). To such economists, even von Wieser's postal service is, at least conceptually, open to such market contestability (although the main example quoted by the new analysts has been that of airlines). The essence of a contestable market is that it is vulnerable to hit-and-run entry: 'Even a very transient profit opportunity need not be neglected by a potential entrant, for he can go in, and before prices change, collect his gains and then depart without cost, should the climate grow hostile' (Baumol 1982, p. 4).

In effect, such new analysis is a theoretical development of the neoclassical concern with perfect competition and especially with its condition of free entry. Indeed, one writer prefers the term 'ultra-free entry' to 'perfect contestability' (Shepherd 1984). What is involved is not only the possibility of a new firm gaining a foothold (which is conventional 'free entry') but the ability to duplicate immediately and entirely replace the existing monopolist. The entrant can, moreover, establish itself before the existing firm makes any price response (the Bertrand–Nash assumption). Finally, exit is perfectly free and without cost. Sunk cost, in other words, is zero. Given these conditions, even the threat of entry (potential competition) may hold price down to cost. A government scheme of regulated prices might therefore be socially detrimental.

Although such theoretical innovation is challenging, it has given rise to considerable controversy concerning both the internal consistency of the theory and empirical support for it. The assumption of zero sunk costs has been the one that has come under most attack. It has been observed for instance that in most markets sunk costs are more obvious in the short run than in the long run; and this is by definition. With *any* element of sunk cost the existing firm has a proportionate potential pricing advantage over an entrant. But it is in the very short period that the pure contestability theory stipulates a zero-price response from the incumbent. Meanwhile, with respect to the question of the empirical basis for the theory, Baumol et al. concede that very little is available so far.

Doubts about the efficiency of government price regulation of natural monopolies have also been raised by Demsetz (1968). He has proposed that formal regulation is unnecessary where governments can allow 'rivalrous competitors' to bid

for the exclusive rights to supply a good or service over a given 'contract period'. The appearance of a single firm may not imply monopoly pricing, because competition could have previously asserted itself at the franchise bidding stage. Monopoly *structure* therefore does not inevitably predict monopoly *behaviour*, although some element of the latter could appear if conditions, say of production, change during the period of the contract.

An ostensibly similar line of argument to that of Demsetz was offered by Bentham and Chadwick. Chadwick's investigation into water supply in London in the 1850s revealed circumstances of natural monopoly. But he argued that inefficiency was prevalent because the field was divided among 'seven separate companies and establishments of which six were originally competing within the field of supply, with two and three sets of pipes down many of the same streets' (quoted in Crain and Ekelund 1976). Following Chadwick's recommendation, rivalry was channelled into what he called competition *for* the field and away from (costly) competition '*within* the field'. The same reasoning applied to the railways. Public ownership was advocated while management (operation) of the services was to be contracted out via a competitive franchise bidding process from among potential private enterprises.

It must next be recognized that very many monopolies, if not most, are *unnatural;* that is, they arise not from inexorable economic conditions but from man- made arrangements, usually through the exercise of political power. In these cases the monopoly is typically awarded by government but not usually with the intention of encouraging the introduction of a new product (as with patents). Instead, one supplier is granted the sole right of trading an existing product or service to the exclusion of all other suppliers. A natural state of competition is thus converted by fiat into one of (statutory) monopoly. In this case the classical analyst might see more potential relevance in Harberger's model of welfare loss from monopoly.

Where the monopoly right is granted by the government, and assuming that price discrimination is prohibitively costly, it would seem, again at first sight, that the monopoly rent or 'prize' to the successful producer could indeed be represented by a rectangle such as *ABCP* in Fig. 1. But since the seminal writing of Tullock (1967), economists have come to recognize that the pursuit of such monopoly rents is itself a competitive activity, and one that consumes resources. Since Krueger (1974) this process has become known as 'rent seeking' and it frequently takes the form of lobbying, offering campaign contributions, bribery, and other ways of influencing the authorities to grant exclusive rights to production, rights that are then policed by the coercive powers of government.

Recent work has modified the conclusion that the value of resources used in pursuit of the rents would exactly equal the value of the rents. Some writers have urged that lobbying by consumers might to some extent offset that of potential monopolists such that a regulated price at a magnitude lower than P (but higher than C) in Fig. 1 would result. In this case, of course, the expected rectangle of monopoly rent would be reduced and the producers collectively would not spend more than this in rent seeking.

Jadlow (1985) has reduced still further the expected magnitude of such monopoly rent rectangles by introducing a multi-period model wherein other rent seekers continue to compete for the valued monopoly prize while consumers, regulators and antitrusters continue their endeavours to eliminate the rents over a protracted period into the future. Since, therefore, instead of a one-time prize, the monopoly rent is viewed as the expected present value of a stream of rents over a series of future time periods in which uncertainty is present, there is likely to be a significant reduction of resources invested in rent-seeking activities.

It is usually implied by economists that the task of public policy with regard to monopoly is to eliminate monopoly profit by one means or another. The above analysis reveals, however, that the conventional measures of social losses via the welfare triangles, plus the rectangles of potential transfers that are partially 'eaten up' by resources devoted to rent-seeking, are

predominantly applicable to monopolies that are politically bestowed. We are thus left with the conclusion that appropriate public policy (according to usual economic reasoning) involves government 'correcting for' something it has created itself. The direct way of solving such a problem, at least to the innocent, would be for the government simply to abstain from granting statutory monopoly privileges in the first place. The newer 'economics of politics', however, has produced reasons why the legislative activity of monopoly rent creation is inherent in the very structure of majority voting democracies. Indeed, some writers (Brennan and Buchanan 1980) argue that the very institution of government is usually a monopoly. In so far as this is true, we face the paradoxical situation that the public policy prescribed in economics textbooks is one whereby monopoly in general is policed or controlled by an institution that is itself a monopoly.

The problem of government sponsored monopolies is currently receiving considerable attention. Indeed, it constitutes one of the most profound issues of the day. For hopeful developments we must look again, presumably, to still further research in the modern economics of politics.

See Also

- ▶ Competition
- ▶ Contestable Markets
- ▶ Market Structure
- ▶ Monopoly

References

Baumol, W.J. 1982. Contestable markets: An uprising in the theory of industry structure. *American Economic Review* 72 (1): 1–15.

Brennan, G., and J. Buchanan. 1980. *The power to tax: Analytical foundations of the fiscal constitution*. Cambridge: Cambridge University Press.

Cowling, K., and D.C. Mueller. 1978. The social costs of monopoly power. *Economic Journal* 88: 727–748.

Crain, W.M., and R.E. Ekelund Jr. 1976. Chadwick and Demsetz on competition and regulation. *Journal of Law and Economics* 19 (1): 149–162.

Demsetz, H. 1968. Why regulate utilities? *Journal of Law and Economics* 11: 55–65.

Demsetz, H. 1982. *Economic, legal, and political dimensions of competition, Professor Dr. F. de Vries Lectures in Economics*. Vol. 4. Amsterdam: North-Holland.

Ely, R.T. 1900. *Monopolies and trusts*. New York: Macmillan.

Fisher, I. 1923. *Elementary principles of economics*. New York: Macmillan.

Gunton, G. 1888. The economic and social aspects of trusts. *Political Science Quarterly* 3 (3): 385–408.

Harberger, A.C. 1954. Monopoly and resource allocation. *American Economic Association, Papers and Proceedings* 44: 77–87.

Jadlow, J.M. 1985. Monopoly rent-seeking under conditions of uncertainty. *Public Choice* 45 (1): 73–87.

Kamerschen, D.R. 1966. An estimation of the 'welfare losses' from monopoly in the American economy. *Western Economic Journal* 4: 221–236.

Krueger, A.O. 1974. The political economy of the rent-seeking society. *American Economic Review* 64: 291–303.

Mill, J.S. 1848. *Principles of political economy,* ed. W.J. Ashley. Reprinted, New York: A.M. Kelley, 1965.

Shepherd, W.G. 1984. 'Contestability' vs. competition. *American Economic Review* 74 (2): 572–587.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. 2 vols, ed. E. Cannan. London: Methuen, 1960.

Stigler, G. 1956. The statistics of monopoly and merger. *Journal of Political Economy* 64: 33–40.

Tullock, G. 1967. The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal* 5: 224–232.

von Wieser, F. 1914. *Social economics*. Trans. A. Ford Hinrichs. New York: A.M. Kelley, 1967.

Worcester, D.A. Jr. 1973. New estimates of the welfare loss to monopoly, United States: 1956–69. *Southern Economic Journal* 40 (2): 234–245.

Monopoly Capitalism

Paul M. Sweezy

Abstract

To Marxian economists, 'monopoly capitalism' denotes the stage of capitalism beginning in the last quarter of the 19th century and maturing after the Second World War. While Marx and Engels wrongly thought it heralded the demise of capitalism, later thinkers, like Sweezy and Baran, have tried to identify its main features

and 'laws of motion'. They argue that, by increasing the savings potential of the economy and reducing opportunities for productive investment, monopoly capitalism suppresses levels of income and employment. No other approach, whether mainstream or traditional Marxian, has satisfactorily explained capitalism's growing tendency towards stagnation in the 20th century.

Keywords

Accumulation of capital; Baran, P.; Burns, A.; Capitalism; Cartellization; Concentration of capital; Engels, F.; Hilferding, R.; Kalecki, M.; Keynesianism; Laws of motion of capitalism; Lenin, V.; Marx, K.; Monopoly capitalism; Stagnation; Sweezy, P.; Veblen, T

JEL Classifications

B1

Among Marxian economists 'monopoly capitalism' is the term widely used to denote the stage of capitalism which dates from approximately the last quarter of the 19th century and reaches full maturity in the period after World War II. Marx's *Capital*, like classical political economy from Adam Smith to John Stuart Mill, was based on the assumption that all commodities are produced by industries consisting of many firms, or capitals in Marx's terminology, each accounting for a negligible fraction of total output and all responding to the price and profit signals generated by impersonal market forces. Unlike the classical economists, however, Marx recognized that such an economy was inherently unstable and impermanent. The way to succeed in a competitive market is to cut costs and expand production, a process which requires incessant accumulation of capital in ever new technological and organizational forms. In Marx's words: 'The battle of competition is fought by cheapening of commodities. The cheapness of commodities depends, *ceteris paribus*, on the productiveness of labour, and this again on the scale of production. Therefore the larger capitals beat the smaller.' Further, the credit system which 'begins as a modest helper of accumulation' soon 'becomes a new and formidable weapon in the competitive struggle, and finally it transforms itself into an immense social mechanism for the centralization of capitals' (Marx 1867, ch. 25, sect. 2). Marx, and even more clearly Engels when preparing the second and third volumes of *Capital* for the printer two decades later, concluded, in the latter's words, that 'the long cherished freedom of competition has reached the end of its tether and is compelled to announce its own palpable bankruptcy' (Marx 1894, ch. 27).

There is thus no doubt that Marx and Engels believed capitalism had reached a turning point. In their view, however, the end of the competitive era marked not the beginning of a new stage of capitalism but rather the beginning of a transition to the new mode of production that would take the place of capitalism. It was only somewhat later, when it became clear that capitalism was far from on its last legs that Marx's followers, recognizing that a new stage had actually arrived, undertook to analyse its main features and what might be implied for capitalism's 'laws of motion'.

The pioneer in this endeavour was the Austrian Marxist Rudolf Hilferding whose magnum opus *Das Finanzkapital* appeared in 1910. A forerunner was the American economist Thorstein Veblen, whose book *The Theory of Business Enterprise* (Veblen 1904) dealt with many of the same problems as Hilferding's: corporation finance, the role of banks in the concentration of capital, etc. Veblen's work, however, was apparently unknown to Hilderding, and neither author had a significant impact on mainstream economic thought in the English-speaking world, where the emergence of corporations and related new forms of business activity and organization, though the subject of a vast descriptive literature, was almost entirely ignored in the dominant neoclassical orthodoxy.

In Marxist circles, however, Hilferding's work was hailed as a breakthrough, and its pre-eminent place in the Marxist tradition was assured when Lenin strongly endorsed it at the beginning of his Imperialism, the Highest Stage of Capitalism. 'In 1910,' Lenin wrote, 'there appeared in Vienna the work of the Austrian Marxist, Rudolf Hilferding, Finance Capital …. This work gives a very

valuable theoretical analysis of "the latest phase of capitalist development", the subtitle of the book.'

As far as economic theory in the narrow sense is concerned, Lenin added little to Finance Capital, and in retrospect it is evident that Hilferding himself was not successful in integrating the new phenomena of capitalist development into the core of Marx's theoretical structure (value, surplus value, and above all the process of capital accumulation). In chapter 15 of his book ('Price Determination in the Capitalist Monopoly. Historical Tendency of Finance Capital') Hilferding, in seeking to deal with some of these problems, came up with a very striking conclusion which has been associated with his name ever since. Prices under conditions of monopoly, he thought, are indeterminate and hence unstable. Wherever concentration enables capitalists to achieve higher than average profits, suppliers and customers are put under pressure to create counter combinations which will enable them to appropriate part of the extra profits for themselves. Thus monopoly spreads in all directions from every point of origin. The question then arises as to the limits of 'cartellization' (the term is used synonymously with monopolization). Hilferding answers:

> The answer to this question must be that there is no absolute limit to cartellization. What exists rather is a tendency to the continuous spread of cartellization. Independent industries, as we have seen, fall more and more under the sway of the cartellized ones, ending up finally by being annexed by the cartellized ones. The result of this process is then a *general cartel*. The entire capitalist production is consciously controlled from one center which determines the amount of production in all its spheres .... It is the consciously controlled society in antagonistic form.

There is more about this vision of a future totally monopolized society, but it need not detain us. Three quarters of a century of monopoly capitalist history has shown that while the tendency to concentration is strong and persistent, it is by no means as ubiquitous and overwhelming as Hilferding imagined. There are powerful countertendencies – the breakup of existing firms and the founding of new ones – which have been strong enough to prevent the formation of anything even remotely approaching Hilferding's general cartel.

The first signs of important new departures in Marxist economic thinking began to appear toward the end of the interwar years, i.e., the 1920s and 1930s; but on the whole this was a period in which Lenin's *Imperialism* was accepted as the last word on monopoly capitalism, and the rigid orthodoxy of Stalinism discouraged attempts to explore changing developments in the structure and functioning of contemporary capitalist economies. Meanwhile, academic economists in the West finally got around to analysing monopolistic and imperfectly competitive markets (especially Edward Chamberlin and Joan Robinson), but for a long time these efforts were confined to the level of individual firms and industries. The so-called Keynesian revolution which transformed macroeconomic theory in the 1930s was largely untouched by these advances in the theory of markets, continuing to rely on the time-honoured assumption of atomistic competition.

The 1940s and 1950s witnessed the emergence of new trends of thought within the general framework of Marxian economics. These had their roots on the one hand in Marx's theory of concentration and centralization which, as we have seen, was further developed by Hilferding and Lenin; and on the other hand in Marx's famous Reproduction Schemes presented and analysed in Volume II of *Capital*, which were the focal point of a prolonged debate on the nature of capitalist crises involving many of the leading Marxist theorists of the period between Engels' death (1895) and World War I. Credit for the first attempt to knot these two strands of thought into an elaborated version of Marxian accumulation theory goes to Michal Kalecki, whose published works in Polish in the early 1930s articulated, according to Joan Robinson and others, the main tenets of the contemporaneous Keynesian 'revolution' in the West. Kalecki had been introduced to economics through the works of Marx and the great Polish Marxist Rosa Luxemburg, and he was consequently free of the inhibitions and preconceptions that went with a training in neoclassical economics. He moved to England in the mid-1930s, entering into the intense discussions and debates of the period and making his own distinctive contributions along the lines of his previous work and that

of Keynes and his followers in Cambridge, Oxford and the London School of Economics. In April 1938 Kalecki published an article in *Econometrica* ('The Distribution of the National Income') which highlighted differences between his approach and that of Keynes, especially with respect to two crucially important and closely related subjects, namely, the class distribution of income and the role of monopoly. With respect to monopoly, Kalecki stated at the end of the article a position which had deep roots in his thinking and would henceforth be central to his theoretical work:

> The results arrived at in this essay have a more general aspect. A world in which the degree of monopoly determines the distribution of the national income is a world far removed from the pattern of free competition. Monopoly appears to be deeply rooted in the nature of the capitalist system: free competition, as an assumption, may be useful in the first stage of certain investigations, but as a description of the normal state of capitalist economy it is merely a myth.

A further step in the direction of integrating the two strands of Marx's thought – concentration and centralization on the one hand and crisis theory on the other – was marked by the publication in 1942 of *The Theory of Capitalist Development* by Paul M. Sweezy, which contained a fairly comprehensive review of the pre-war history of Marxist economics and at the same time made explanatory use of concepts introduced into mainstream monopoly and oligopoly theory during the preceding decade. This book, soon translated into several foreign languages, had a significant effect in systematizing the study and interpretation of Marxian economic theories.

It should not be supposed, however, that these new departures were altogether a matter of theoretical speculation. Of equal if not greater importance were the changes in the structure and functioning of capitalism which had emerged during the 1920s and 1930s. On the one hand the decline in competition which began in the late 19th century proceeded at an accelerated pace – as chronicled in the classic study by Arthur R. Burns, *The Decline of Competition: A Study of the Evolution of American Industry* (1936) – and on the other hand the unprecedented severity of the depression of the 1930s provided dramatic proof of the inadequacy of conventional business cycle theories. The Keynesian revolution was a partial answer to this challenge, but the renewed upsurge of the advanced capitalist economies during and after the war cut short further development of critical analysis among mainstream economists, and it was left to the Marxists to carry on along the lines that had been pioneered by Kalecki before the war.

Kalecki spent the war years at the Oxford Institute of Statistics whose Director, A. L. Bowley, had brought together a distinguished group of scholars, most of them emigrés from occupied Europe. Among the latter was Josef Steindl, a young Austrian economist who came under the influence of Kalecki and followed in his footsteps. Later on, Steindl recounted the following:

> On one occasion I talked with Kalecki about the crisis of capitalism. We both, as well as most socialists, took it for granted that capitalism was threatened by a crisis of existence, and we regarded the stagnation of the 1930s as a symptom of such a major crisis. But Kalecki found the reasons, given by Marx, why such a crisis should develop, unconvincing; at the same time he did not have an explanation of his own. I still do not know, he said, why there should be a crisis of capitalism, and he added: Could it have anything to do with monopoly? He subsequently suggested to me and to the Institute, before he left England, that I should work on this problem. It was a very Marxian problem, but my methods of dealing with it were Kaleckian (Steindl 1985).

Steindl's work on this subject was completed in 1949 and published in 1952 under the title *Maturity and Stagnation in American Capitalism*. While little noticed by the economics profession at the time of its publication, this book nevertheless provided a crucial link between the experiences, empirical as well as theoretical, of the 1930s, and the development of a relatively rounded theory of monopoly capitalism in the 1950s and 1960s, a process which received renewed impetus from the return of stagnation to American (and global) capitalism during the 1970s and 1980s.

The next major work in the direct line from Marx through Kalecki and Steindl was Paul Baran's book, *The Political Economy of Growth*, which presented a theory of the dynamics of

monopoly capitalism and opened up a new perspective on the nature of the interaction between developed and underdeveloped capitalist societies. This was followed by the joint work of Baran and Sweezy, *Monopoly Capital: An Essay on the American Economic and Social Order*, incorporating ideas from both of their earlier works and attempting to elucidate, in the words of their Introduction, the 'mechanism linking the foundation of society (under monopoly capitalism) with what Marxists call its political, cultural, and ideological superstructure'. Their effort, however, still fell short of a comprehensive theory of monopoly capitalism since it neglected 'a subject which occupies a central place in Marx's study of capitalism', that is, a systematic inquiry into 'the consequences which the particular kinds of technological change characteristic of the monopoly capitalist period have had for the nature of work, the composition (and differentiation) of the working class, the psychology of workers, the forms of working-class organization and struggle, and so on.' A pioneering effort to fill this gap in the theory of monopoly capitalism was taken by Harry Braverman a few years later (Braverman 1974) which in turn did much to stimulate renewed research into changing trends in work processes and labour relations in the late 20th century.

Marx wrote in the Preface to the first edition of volume 1 of *Capital* that 'it is the ultimate aim of this work to lay bare the economic law of motion of modern society'. What emerged, running like a red thread through the whole work, could perhaps better be called a theory of the accumulation of capital. In what respect, if at all, can it be said that latter-day theories of monopoly capitalism modify or add to Marx's analysis of the accumulation process?

As far as form is concerned, the theory remains basically unchanged, and modifications in content are in the direction of putting even greater emphasis on certain tendencies already demonstrated by Marx to be inherent in the accumulation process. This is true of concentration and centralization, and even more spectacularly so of the role of what Marx called the credit system, now grown to monstrous proportions compared to the small beginnings of his day. In addition, and perhaps most important, the new theories seek to demonstrate that monopoly capitalism is more prone than its competitive predecessor to generating unsustainable rates of accumulation, leading to crises, depressions, and prolonged periods of stagnation.

The reasoning here follows a line of thought which recurs in Marx's writings, especially in the unfinished later volumes of *Capital* (including *Theories of Surplus Value*): individual capitalists always strive to increase their accumulation to the maximum extent possible and without regard for the ultimate overall effect on the demand for the increasing output of the economy's expanding capacity to produce. Marx summed this up in the well-known formula that 'the real barrier of capitalist production is capital itself'. The upshot of the new theories is that the widespread introduction of monopoly raises this barrier still higher. It does this in three ways.

1. Monopolistic organization gives capital an advantage in its struggle with labour, hence tends to raise the rate of surplus value and to make possible a higher rate of accumulation.
2. With monopoly (or oligopoly) prices replacing competitive prices, a uniform rate of profit gives way to a hierarchy of profit rates – highest in the most concentrated industries, lowest in the most competitive. This means that the distribution of surplus value is skewed in favour of the larger units of capital which characteristically accumulate a greater proportion of their profits than smaller units of capital, once again making possible a higher rate of accumulation.
3. On the demand side of the accumulation equation, monopolistic industries adopt a policy of slowing down and carefully regulating the expansion of productive capacity in order to maintain their higher rates of profit.

Translated into the language of Keynesian macro theory, these consequences of monopoly mean that the savings potential of the system is increased, while the opportunities for profitable investment are reduced. Other things being equal,

therefore, the level of income and employment under monopoly capitalism is lower than it would be in a more competitive environment.

To convert this insight into a dynamic theory, it is necessary to see monopolization (the concentration and centralization of capital) as an ongoing historical process. At the beginning of the transition from the competitive to the monopolistic stage, the accumulation process is only minimally affected. But with the passage of time the impact grows and tends sooner or later to become a crucial factor in the functioning of the system. This, according to monopoly capitalist theory, accounts for the prolonged stagnation of the 1930s as well as for the return of stagnation in the 1970s and 1980s following the exhaustion of the long boom caused by World War II and its multifaceted aftermath effects.

Neither mainstream economics nor traditional Marxian theory had been able to offer a satisfactory explanation of the stagnation phenomenon which has loomed increasingly large in the history of the capitalist world during the 20th century. It is thus the distinctive contribution of monopoly capitalist theory to have tackled this problem head on and in the process to have generated a rich body of literature which draws on and adds to the work of the great economic thinkers of the last 150 years. A representative sampling of this literature, together with editorial introductions and interpretations, is contained in Foster and Szlajfer (1984).

See Also

- ▶ Baran, Paul Alexander (1910–1964)
- ▶ Capitalism
- ▶ Finance
- ▶ Foreign Direct Investment
- ▶ Marx, Karl Heinrich (1818–1883)

Bibliography

Baran, P.A. 1957. *The political economy of growth*. New York: Monthly Review Press.

Baran, P.A., and P.M. Sweezy. 1966. *Monopoly capital: An essay on the American economic and social order*. New York: Monthly Review Press.

Braverman, H. 1974. *Labor and monopoly capital: The degradation of work in the twentieth century*. New York: Monthly Review Press.

Burns, A.R. 1936. *The decline of competition: A study of the evolution of American industry*. New York: McGraw-Hill.

Foster, J.B., and H. Szlajfer, eds. 1984. *The faltering economy: The problem of accumulation under monopoly capitalism*. New York: Monthly Review Press.

Hilferding, R. 1910. *Das Finanzkapital*. Trans. M. Watnick and S. Gordon as *Finance capital*, ed. T. Bottomore. London: Routledge & Kegan Paul, 1981.

Kalecki, M. 1938. The distribution of the national income. *Econometrica*.

Lenin, V.I. 1917. *Imperialism, the highest stage of capitalism*.

Marx, K. 1867. *Capital*, vol. 1. Moscow: Progress Publishers.

Marx, K. 1885. *Capital*, vol. 2. Moscow: Progress Publishers.

Marx, K. 1894. *Capital*, vol. 3. Moscow: Progress Publishers.

Steindl, J. 1952. *Maturity and stagnation in American capitalism*. Oxford: Blackwell.

Steindl, J. 1985. The present state of economics. *Monthly Review*.

Sweezy, P.M. 1942. *The theory of capitalist development*. New York: Monthly Review Press.

Sweezy, P.M. 1966. See Baran and Sweezy (1966).

Veblen, T. 1904. *The theory of business enterprise*. New York: Charles Scribner's Sons.

Monopsonistic Discrimination and the Gender Wage Gap

Erling Barth

Abstract

Monopsonistic discrimination refers to a situation in which employers differentiate pay between groups of workers who exhibit different elasticities of labour supply. The concept of dynamic monopsony has revived the idea of monopsonistic discrimination in the labour market. As there are frictions in the job-to-job mobility of workers, firms may exercise market power even in labour markets with thousands of employers. If there are more frictions in the labour market for women than for men, a gender wage gap may arise as employers exploit this difference and segment their pay policy towards each gender.

Keywords
Discrimination; Gender pay gap; Job mobility; Labour markets; Labour supply; Monopsony; Wage determination; Wage dispersion

JEL Classifications
D43; J3; J43; J63; J71

Introduction

Monopsonistic discrimination in the labour market refers to a situation in which employers differentiate wages between groups of employees because the employers possess different degrees of market power towards each of those groups of employees. Market power may be measured by the elasticity of labour supply with respect to wages. What percentage of the workers will the employer lose by cutting wages by 1%? If this elasticity is small for a particular group, the employer may exercise market power towards that group; if the elasticity is large, the employer possesses little market power and has to pay a market wage to retain its workers. In a labour market with many frictions, the elasticity of supply is small, and employers' market power is large.

The gender wage gap refers to the fact that, on average, women have lower wages than men, and to the fact that female-dominated jobs and occupations tend to pay less than male-dominated jobs and occupations. Even though the gender wage gap has declined over the last century (see Palgrave: Blau and Kahn) the remaining difference has been remarkably persistent across countries and over time during recent decades (see Blau and Kahn (2006) for US evidence). This is particularly puzzling, considering the 'quiet revolution' of the last century, as described by Claudia Goldin (2006), with a large increase in women's labour market participation, and women catching up with men in terms of their human capital in many countries.

The concept of monopsonistic discrimination offers a potential explanation for the existence and persistence of the gender wage gap. The idea is simple: a monopsonist, originally a term referring to a single buyer in a market (see Manning 2003), sets wages below marginal revenue product (see Palgrave: Manning). The more inelastic the labour supply, the lower are the wages relative to productivity, simply because an employer facing inelastic labour supply loses a smaller share of its workers by cutting wages than an employer facing more elastic labour supply. By differentiating wages between groups of workers with different elasticities of labour supply, the monopsonist may thus obtain higher profits, in the same way that an airline may obtain higher profits from differentiating fares between groups of customers with different demand elasticities. If female labour supply is more inelastic than male labour supply, women will earn less than men relative to their productivity.

Since men and women conveniently sort into different occupations, such differentiation may be conducted without breaking anti-discrimination laws, and may even be self-enforcing by inducing both increased gender segregation across occupation and increased differences in behaviour between men and women.

This theory of the gender wage gap, first proposed by Joan Robinson (1933), eventually lost its credibility as an explanation. Recently, however, it has received renewed interest, in particular through a renewal of the concept of monopsony in the labour market (see Palgrave: Manning). This article explains the theory of monopsonistic discrimination as a mechanism behind the gender wage gap, and lays out how this theory regained interest and is now seen as a potentially powerful explanation of the gender wage gap in modern labour markets.

The Roots

Joan Robinson (1933) was the first to develop the concept of monopsonistic discrimination. She modelled the behaviour of an employer who is the only buyer, i.e. a monopsonist, in the labour market. Think of the large employer in a company town, or the public health service in a community.

Whereas an employer in an idealised competitive labour market without frictions faces a horizontal labour supply curve, the monopsonist faces an upward sloping labour supply curve, and may set its wages unilaterally. If it pays more, it may employ more workers. If it pays less it may employ fewer workers. However, since workers have nowhere else to go, it does not lose all its workers by paying less. Thus, it pays to lower the wages if the loss in terms of workforce is sufficiently small compared to the gain from cutting wages to all. It turns out that the optimal wage-setting behaviour of the monopsonist is to pay the least to the group of workers whose labour supply is the least elastic (see below). Intuitively, if the supply curve is steep, it does not cost much in terms of employment to cut wages, whereas if it is flat, one loses a lot of workers by paying less.

The theory explains why employers may gain from differentiating wages, and if female labour supply is less elastic than male labour supply, it may explain the gender wage gap.

Over the years, this theory was basically discarded as not very relevant for two good reasons: situations with a pure monopsony seem increasingly rare, and women's labour supply on the margin between working and not working appears to be more, rather than less, wage elastic than men's labour supply. So the theory seemed to be only that – a theory. Recently, however, the theory of monopsonistic behaviour in the labour market has gained a lot of attention, and is now applied to labour markets with thousands of employers. It also seems much more reasonable to expect that female labour supply may be less elastic than men's. How did this change come about?

The Modern Theory of Dynamic Monopsony

The modern theory of monopsonistic wage setting departs from the observation that there are *frictions* in almost any labour market (see Palgrave: Manning). There is lack of information about jobs, about job attributes, about workers, and about worker attributes; furthermore, there are establishment-specific skills, personal ties and relationships, and location specific ties or preferences, and other factors that limit the speed with which markets may adjust. With frictions, search theory becomes an important tool with which to understand market behaviour.

Within the framework of equilibrium search, work in particular by the Nobel Prize winner Dale Mortensen (Burdett and Mortensen 1998), revived the theory of monopsonistic behaviour in the labour market. The newer version of monopsony is labelled dynamic monopsony because it arises from the firms' dynamic problem of maintaining a given workforce over time when facing restrictions both in terms of hiring and in retaining workers (Manning 2003). Burdett and Mortensen (1998) model firms' behaviour in a labour market in which workers are engaged in job-to-job search. Consider a labour market where, because of frictions, firms may hire from both the pool of unemployment and also from all workers employed in other firms, and the number of employees an employer may hire in a given period of time is a continuous and increasing function of the wage $H(w)$, $H'(w) > 0$. In the benchmark competitive case, $H'(w)$ is infinite, and an employer cannot hire any workers if it pays one cent below the competitive wage, but may hire any number of workers by paying the competitive wage.

At the same time, the share of the stock of employees leaving the firm is a decreasing function of the wage $q(w)$, $q'(w) < 0$. Some workers leave for exogenous reasons, but others leave to get higher paying jobs elsewhere. If the wage is high, fewer workers quit to higher paying jobs. Quits are assumed proportional to the number of employees, while the number of hires is independent of L (the current stock of employees).

In steady state the number of hires for any one firm has to equal its number of quits, $H(w) = Lq(w)$ and the steady state labour supply to a firm is thus given by $L(w) = H(w)/q(w)$, with $L'(w) > 0$. In the benchmark competitive case, $L(w)$ is horizontal and $L'(w)$ is infinite. In the more realistic case with some frictions, $L(w)$ is an upward sloping supply curve facing each employer, where the slope is determined by the level of frictions.

Intuitively, since a given share of workers quit every period, a firm that wants to maintain a large stock of employees has to hire more workers every period than a firm that wants to maintain a smaller stock of employees. By posting a higher wage, a firm both increases hires and decreases the quit rate, and is thus able to sustain a higher level of employment. Notice that when there are frictions in the labour market, it may be more important for the firm to attract workers from other firms than to attract workers from the pool of unemployed. Likewise, it may be more important to keep workers from moving to other firms than to keep workers from exiting from the labour market. The margin of job-to-job mobility may be the dominant margin when it comes to maintaining a given stock of workers. This situation is relevant for most labour markets regardless of size, which is one reason why monopsonistic wage setting is now seen as highly relevant, even in labour markets with thousands of employers.

Behaving as a monopsonistic wage setter, a profit-maximising employer chooses a wage according to the first-order condition $(p - w)\frac{\partial L}{\partial w} - L = 0$, where p is the marginal revenue product, implying that the wage may be written as

$$p = (1 + 1/\varepsilon)w = \omega w$$

where the *mark-up*, ω, depends negatively on the elasticity of labour supply, ε, facing the firm (see Palgrave: Manning). If the elasticity of labour supply is small, indicating large frictions, the mark-up is large, and wages are lower relative to productivity. The first key result is thus that the wage level depends positively on the elasticity of labour supply facing each firm.

An Equilibrium Wage Distribution

The Burdett–Mortensen model provides an equilibrium of wages and employment across firms, and the second key result is that even in a world with homogeneous productivity and workers, different firms offer different wage levels. A free entry condition ensures that large firms earn the same overall profit as small firms, and optimal behaviour on the part of each firm is only compatible with a continuous wage distribution across firms. In equilibrium, a large firm offers a higher wage and employs more workers, but makes less profit per worker, while the small firm offers a lower wage, employs fewer workers, but makes more profit per worker. So even for identical workers, different firms offer different wages. This provides workers with incentives to search for better jobs in other firms. Since there are frictions, workers in lower paying firms cannot choose to leave for a better paying firm before they get an offer. In this sense, they have both to find and queue up for jobs in higher paying firms.

In the more realistic case with productivity differences across firms, frictions in the labour market again create a wage distribution across firms, even for homogeneous workers. In this case, wages tend to grow with productivity as the more productive firms choose to be bigger; as a result of wanting to maintain a larger workforce, they have to pay higher wages. For workers, wages turn out to be a weighted sum of their reservation wage and the marginal product. The bottom of the wage distribution is given by the reservation wage and the top is given by a weighted sum of the reservation wage and the marginal revenue product, with weights being determined by the level of frictions in the labour market.

The elasticity of labour supply towards any firm thus depends crucially on the distribution of wages and employment surrounding it, on the search efforts of its own workers, the search effort of other firms and the search efforts of the workers in other firms.

Monopsony and the Gender Pay Gap

Barth and Dale-Olsen (2009) apply the model of monopsonistic wage setting to the analysis of the gender wage gap. The profit of an employer who employs two types of workers is given by $R[L1(w1), L2(w2)] - w1L1(w1) - w2L2(w1)$, where $R[\cdot]$ is revenue. Assuming that each labour input is independent in production and has the same marginal revenue product, p, the two first-

order conditions for a wage-setting employer, laid out above, imply that the pay gap between the two types of workers may be written as:

$$\frac{w^2 - w^1}{w^1} = \frac{\omega^1}{\omega^2} - 1 = \frac{\varepsilon^2 - \varepsilon^1}{\varepsilon^1(\varepsilon^2 + 1)} \qquad (1)$$

The wage gap is increasing in the relative mark-up between the two groups. If the mark-up for group 1 is larger than the mark-up for group 2, the wage gap is positive. As noted above, the mark-up is larger when the elasticity of labour supply is smaller. The last equation reflects this point, where we see that the wage gap is increasing in the elasticity of group 2 relative to the elasticity of labour supply of group 1. The firm, in this case, exploits the fact that the two groups of workers have different elasticities of supply, and maximise profits by differentiating wages. The key ingredient is the difference in frictions in the labour market. If women have less elastic labour supply, the employer may reduce wages for women without losing as many workers as they would if they reduced the wages of men. What then, could be the reasons for gender differences in the elasticity of labour supply facing each firm? Why do women's job-to-job transitions appear to be less sensitive to wages than men's job-to-job transitions appear to be?

One factor that seems likely to be at play is different preferences between pecuniary and non-pecuniary features of jobs. Given the traditional division of labour at home, women may be more attracted to jobs that offer possibilities of combining care obligations and work. Put differently, they may be more attracted to jobs or occupations that do not punish such obligations. Men and women may also have different preferences towards different types of reward structures. Many care work occupations are typically characterised by flat wage profiles, whereas many male-dominated occupations are typically characterised by very steep wage profiles. In an analysis of the relationship between household allocation of tasks and the rewards to effort in the labour market, Albanesi and Olivetti (2009) model endogenous selection between men and women by pay for performance schemes. Goldin (2014) shows how convex reward structures in firms, particularly rewarding long hours or particular hours, work to sustain the gender gap in pay.

If women care more about non-pecuniary features of jobs, they may also be more willing to take a wage cut rather than moving to another employer. The point here is not that some people may be willing to pay a compensating differential to have a more suited job, but rather that the non-pecuniary aspects of jobs may make alternatives less attractive or more difficult to come by, and thus be a hindrance to search behaviour and job-to-job transitions. Lower search and less wage-sensitive mobility may be exploited by the employer to pay even less. A related argument comes from the idea that women appear to choose between fewer jobs, and thus have smaller market opportunities than men, as per the 'overcrowding' idea of Bergmann (1974). Also, job-to-job changes may involve higher costs in periods with small children or care obligations for older parents, and may as such in themselves be seen as an activity involving convex rewards to effort and time use.

Differential access to different jobs, such as hiring discrimination in certain sectors or occupations, entry barriers into typical male-dominated occupations or a glass ceiling preventing women from accessing the top-level jobs in each occupation, produces a wedge between men's and women's incentives to search for better jobs. Again, the result will be a less elastic labour supply on the part of women. Hiring discrimination in employment may thus be a factor that fuels monopsonistic wage discrimination.

Discrimination

Monopsonistic discrimination is distinct from other types of discrimination in the labour market. While taste-based discrimination (Palgrave: Charles and Guryan; Becker 1971) is based on prejudice and individual preferences among employers, and statistical discrimination (see Palgrave: Moro; Arrow 1973) is based on stereotypes arising from the use of group-based statistics, monopsonistic discrimination between groups is based on differential market power towards these

groups. The employer maximises profits by differentiating wages between different segments of its workforce. In that sense, monopsonistic discrimination is conceptually more related to price discrimination (see Palgrave: Miravete), where a monopolist seller chooses different prices for different market segments. Analogous with the above discussion of employment discrimination, both taste-based and statistical discrimination may serve to reduce women's incentives and opportunities for job-to-job changes, and thus tend to reinforce the incentives for employers to exercise their market power and engage in monopsonistic discrimination.

Historically it was common practice to offer different wages for men and women, even when performing the same tasks for the same employer. A justification of this practice was men's role as breadwinners for the family, while women's paid work was regarded as supplementary income for the household. This practice was, however, outlawed in most countries during the 1960s and 1970s, and discrimination based on individual characteristics such as gender and race is now illegal (see Palgrave: Donohue III) in most countries. How is it possible, then, that employers segment the labour market between men and women in order to offer different wages?

One answer to this question lies in the prevalence and magnitude of gender segregation in the labour market. Women and men choose different educations, choose different occupations and face different access to different types of jobs. Employers in different industries may thus choose different wage policies, and by differentiating between types of occupations or by different jobs and tasks within occupations, employers may effectively differentiate between men and women, without doing so directly based on individual characteristics.

Empirical Evidence

Green et al. (1996) presented early evidence in support of monopsonistic discrimination by identifying larger size wage effects for women than for men, an observation which is consistent with a model of monopsonistic discrimination in the labour market. Manning (1996) analyses relative female employment following from the large rise in the relative earnings of women in the UK after the Equal Pay act of 1970 was passed. He attributes the observation that relative female employment did not fall to monopsony in the female labour market. Using data on high school and college graduates, Bowlus (1997) identifies higher labour market frictions for women than men. Her study was the first to apply an equilibrium search model to gender wage differentials. Bowlus finds that the differences in search parameters explain 20–30% of overall male–female wage differentials of high school and college graduates.

Barth and Dale-Olsen (2009) use the relationship between turnover and the elasticity of labour supply, and utilise linked employer–employee data to analyse the relationship between worker turnover and wages separately for each gender. They find that the turnover of women is less wage elastic than the turnover of men, and thus that employers have an incentive to engage in monopsonistic discrimination in order to increase their profit. More recent evidence is provided by Hirsch et al. (2010), Webber (2013) and Ransom and Oaxaca (2010). An application to the immigrant wage gap is provided by Nanos and Schluter (2014). Recent evidence of monopsony in the labour market includes Staiger et al. (2010) and Falch (2010), who use quasi-experimental designs to identify the elasticity of labour supply.

In addition to empirical studies focusing on monopsony or monopsonistic discrimination, there is a host of evidence that provides support to important features or implications of the model. First of all, there is clear evidence that female-dominated occupations pay less than male-dominated occupations (see, for instance, Lucifora and Reilly (1990)). There is also evidence of wage distribution across firms and establishments – even among homogeneous workers, one of the key implications of the dynamic monopsony model. Barth et al. (2015)

show that a large part of the widening of the US wage distribution may be attributed to increasing wage differentials across establishments. Card et al. (2015) show how gender differences arise from wage differences across establishments, and that women obtain a smaller share of the wage premium across establishments than men.

There is also evidence on the differences in mobility and search behaviour between men and women that may explain differences in the elasticity of labour supply facing each firm (Manning 2003). Early examples are Loprest (1993), who found lower returns to mobility for younger women than for younger men, Sicherman (1996), who found that a larger proportion of women than men leave a job for non-market reasons, and Keith and McWilliams (1999), who found that women do less job-related search than men.

Summing Up

Monopsonistic discrimination occurs when employers exploit differences in their market power towards certain groups by offering different levels of pay. If the supply of female labour is less wage elastic than the supply of male labour, a gender wage gap may arise from this differential treatment. The model of monopsonistic discrimination combines both demand and supply side mechanisms to explain how systematic wage differences may occur. Its practice would be reinforced by the high level of gender segregation in the labour market and it would lead to more compressed wages among women than among men.

This theory is complementary, rather than an alternative, to many of the important explanations of the gender wage gap. For example, hiring discrimination in the labour market may have a direct effect on the wages of men and women, but by limiting job options for women it also reinforces the differences in labour market frictions that underlie monopsonistic discrimination. Also, the convex reward structures in many firms and industries, highlighted by Goldin (2014) as a key source of the remaining pay gap between men and women, may increase the segregation of men and women into different occupations and increase frictions.

One of the reasons why the remaining gender wage gap may be difficult to combat is that it is supported by behaviour, both by employers in the labour market and by workers in households, in ways that tend to reinforce each other. Monopsonistic wage discrimination may thus be one of the key ingredients behind an apparent persistence of the gender wage gap.

See Also

- ▶ Anti-Discrimination Law
- ▶ Feminist Economics
- ▶ Gender Differences (Experimental Evidence)
- ▶ Gender Roles and Division of Labour
- ▶ Inequality (International Evidence)
- ▶ Labour Economics
- ▶ Labour Economics (New Perspectives)
- ▶ Labour Supply
- ▶ Monopsony
- ▶ Mortensen, Dale T. (Born 1939)
- ▶ Price Discrimination (Theory)
- ▶ Robinson, Joan Violet (1903–1983)
- ▶ Statistical Discrimination
- ▶ Taste-Based Discrimination
- ▶ Women's Work and Wages

Bibliography

Albanesi, S., and C. Olivetti. 2009. Home production, market production and the gender wage gap: Incentives and expectations. *Review of Economic Dynamics* 12(1): 80–107.

Arrow, K.J. 1973. The theory of discrimination. In *Discrimination in labour markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.

Barth, E., and H. Dale-Olsen. 2009. Monopsonistic discrimination, worker turnover and the gender wage gap. *Labour Economics* 16: 589–597.

Barth, E., A. Bryson, J.C. Davis, and R.B. Freeman. 2015. It's where you work: Increases in earnings dispersion across establishments and individuals in the U.S. *Journal of Labor Economics* (forthcoming).

Becker, G.S. 1971. *The economics of discrimination*. 2nd ed. Chicago: Chicago University Press.

Bergmann, B.R. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* 1(2): 103–110.

Blau, F.D., and L.M. Kahn. 2006. The U.S. gender pay gap in the 1990s: Slowing convergence. *Industrial and Labour Relations Review* 60(1): 45–66.
Burdett, K., and D. Mortensen. 1998. Equilibrium wage differentials and employer size. *International Economic Review* 39: 257–273.
Card, D., A.R. Cardoso, and P. Kline. 2015. *Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women*, *NBER Working Papers 21403*. Cambridge: National Bureau of Economic Research.
Falch, T. 2010. Teacher mobility responses to wage changes: Evidence from a quasi-natural experiment. *American Economic Review* 101(3): 460–465.
Goldin, C. 2006. The quiet revolution that transformed women's employment, education, and family. *American Economic Review* 96(2): 1–21.
Goldin, C. 2014. A grand gender convergence: Its last chapter. *American Economic Review* 104(4): 1091–1119.
Green, F., S. Machin, and A. Manning. 1996. The employer size-wage effect: Can dynamic monopsony provide an explanation? *Oxford Economic Papers* 48: 433–455.
Hirsch, B., T. Schank, and C. Schnabel. 2010. Differences in labour supply to monopsonistic firms and the gender pay gap: An empirical analysis using linked employer-employee data from Germany. *Journal of Labour Economics* 28(2): 291–330.
Keith, K., and A. McWilliams. 1999. The returns to mobility and job search by gender. *Industrial and Labour Relations Review* 52: 463–474.
Loprest, P.J. 1993. Gender differences in wage growth and job mobility. *American Economic Review* 82: 526–532.
Lucifora, C., and B. Reilly. 1990. Wage discrimination and female occupational intensity. *Labour* 4: 147–168.
Manning, A. 1996. The Equal Pay Act as an experiment to test theories of the labour market. *Economica* 63: 191–212.
Manning, A. 2003. *Monopsony in motion. Imperfect competition in labour markets*. Princeton: Princeton University Press.
Nanos, P., and C. Schluter. 2014. The composition of wage differentials between migrants and natives. *European Economic Review* 65(C): 23–44.
Ransom, M., and R. Oaxaca. 2010. New market power models and sex differences in pay. *Journal of Labour Economics* 28(2): 267–290.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Sicherman, N. 1996. Gender differences in departures from a large firm. *Industrial and Labour Relations Review* 49: 484–505.
Staiger, D.O., J. Spetz, and C.S. Phibbs. 2010. Is there monopsony in the labour market? Evidence from a natural experiment. *Journal of Labour Economics* 28(2): 211–236.
Webber, D.A. 2013. *Firm market power and the earnings distribution*, *IZA Discussion Papers 7342*. Bonn: Institute for the Study of Labor (IZA).

Monopsony

Alan Manning

Abstract

Monopsony refers to the situation where a firm has some market power over the price it pays for its inputs, so that a higher price must be paid the more input is used. Monopsony could exist in any input market but is usually discussed in the context of the labour market. Employers will have monopsony power over their workers because of frictions in the labour market. Employers will use this monopsony power to pay workers less than their marginal product. This gap between marginal product and wage offers policy an opportunity to raise the wage of workers without necessarily jeopardizing their employment.

Keywords

Collusion; Frictions; Human capital; Job mobility; Labour markets; Labour supply; Law of one wage; Minimum wage; Monopsony; Oligopsony; Partial equilibrium; Robinson, J.; Search capital; Search models; Smith, A.; Training; Wage discrimination; Wage dispersion; Wage heterogeneity, sources of

JEL Classifications

D43

The definition of a monopsony in the *Oxford English Dictionary* (OED) is 'a market situation in which there is only one buyer'. Joan Robinson (1933) is credited with inventing the term (but see Thornton 2004, for a discussion of the origins of the term) as a counterpart to the more commonly used and understood term 'monopoly'.

Taken literally, it is very likely that a pure monopsony has never existed in any market, but the term is more generally used to denote a situation in which the supply curve to an individual firm has an input price elasticity that is finite, that

is, is increasing in the input price, and this article follows that usage. If one is pedantic, one might think that 'oligopsony' is a more accurate term to use (defined by the OED as 'a state of the market in which only a small number of buyers exists for a product'), or 'oligopsonistic competition' if one believes that free entry of firms will bid away any monopsony rents.

The market for any type of good or service could, in principle, be monopsonistic. To give some examples from the economic literature, Schroeter (1988) considers the meat-packing industry as an oligopsonistic buyer of cattle, Just and Chern (1980) consider the tomato-canning industry as an oligopsonistic buyer of tomatoes, and Murray (1995) considers saw-mills as oligopsonistic buyers of logs. But the idea of monopsony is most commonly applied to the labour market, and this article focuses on that application. Employers are often felt to have monopsony power only in a few specific labour markets – those for professional athletes in the United States, nurses and teachers (for whom outside cities there may only be one potential employer), and miners and mill workers in company towns in the early days of the Industrial Revolution are some of the more common examples. But, in recent years, some labour economists have argued that monopsony is pervasive in all labour markets.

The plan of this article is the following. We first review the simple partial equilibrium of monopsony, discussing the differences from and similarities to the more conventional perfectly competitive model. We then discuss why it is plausible to believe that employers have some monopsony power over their workers, after which we discuss how the monopsony perspective can help us to a better understanding of the workings of labour markets. The monopsonistic approach is more in line with the way that workers and employers experience the labour market, and can explain a wide range of what are puzzles and anomalies from the perspective of labour markets as perfectly competitive. Many of these puzzles and anomalies have other potential explanations but monopsony offers a simple unified account of their existence.

The Simple Textbook Model of Monopsony

In a perfectly competitive labour market, an employer can hire as many workers of a particular type as it wants at the market wage for that type of workers (and none at all if it tries to pay below the market wage). But, if an employer has some Monopsony power, the labour supply to an individual employer depends positively on the wage paid. The wage elasticity of the labour supply curve facing the firm is therefore finite not infinite. Figure 1 represents such a labour supply curve.

How does this affect the decisions of employers? Denote the supply of labour to the firm if it pays w by $N(w)$ and the inverse of this relationship by $w(N)$. Total labour costs are given by $w(N)N$. Assume that the firm has a revenue function $Y(N)$ and is a simple monopsonist who has to pay a single wage to all its workers. It wants to choose N to maximize profits which are given by:

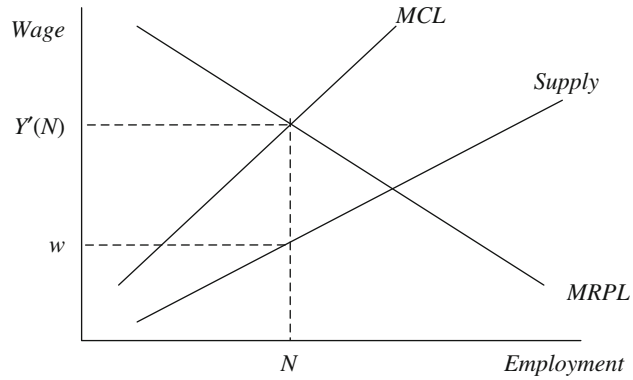
$$\pi = Y(N) - w(N)N. \tag{1}$$

This leads to the first-order condition:

$$Y'(N) = w(N) + w'(N)N. \tag{2}$$

The left-hand side of (2) is the marginal revenue product of labour. The right-hand side is the marginal cost of labour, the increase in total labour costs when an extra worker is hired. The marginal cost of labour (MCL) has two parts: the wage, w, that must be paid to the new worker hired and the increase in wages that must be paid to all existing workers. The MCL is always above the labour supply curve to the firm and is also drawn on Fig. 1. The profit maximizing employer will choose the level of employment where MRPL = MCL and the wage necessary to supply this amount of labour – the solution is represented graphically in Fig. 1.

In equilibrium, the wage paid to workers is less than their marginal revenue product. Although the employer is making positive profit on the marginal worker they have no incentive to increase employment because doing so would require



Monopsony, Fig. 1 The textbook model of monopsony

increasing the wage (to attract the extra worker) and this higher wage must be paid not just to the new worker but also to all the existing workers. One particularly useful way of representing the choice of the firm is that marginal cost of labour is a mark-up on the wage, the mark-up being given by the elasticity of the labour supply curve facing the firm. Write the elasticity of the labour supply curve facing the firm as $\varepsilon_{Nw} = \frac{wN'(w)}{N(w)}$ and let ε be the inverse of this elasticity. Then (2) can be written as:

$$\frac{Y' - w}{w} = \frac{1}{\varepsilon_{Nw}} = \varepsilon \tag{3}$$

so that the proportional gap between the wage and the marginal revenue product is a function of the elasticity of the labour supply curve facing the firm. Perfect competition can be thought of as a special case of this model where $\varepsilon_{Nw} = \infty$ and $\varepsilon = 0$, in which case (3) says that the wage will be equal to the marginal revenue product.

Some of the comparative statics of the monopsony model are the same as the perfectly competitive model and some are different. For example, consider an increase in the marginal revenue product of labour for a single firm – this will lead to an increase in employment and a rise in wages in a monopsony model. The former would occur in a competitive model but the latter would not, as a competitive firm would simply continue to pay the market wage (which would not change if the change in the MRPL affected only a single firm). The impact of shifts in the labour supply curve to the firm is more complicated as the impact depends on how the change affects the marginal cost of labour and not just the average cost of labour. An increase in the supply of labour to the firm that keeps the elasticity the same will result in a rise in employment and a fall in wages, just as in the competitive model. But matters are more complicated if the elasticity of the labour supply curve can also change as the average and marginal cost of labour can move in opposite directions, the most familiar example of which is the impact of a minimum wage. The minimum wage raises the average cost of labour but (if it is binding) reduces $w'(N)$ so its effect on the marginal cost of labour (see (2)) is ambiguous. In fact, one can show that a minimum wage that just binds must raise employment (a demonstration of this can be found in most labour economics textbooks).

Although the model described here captures the fundamentals of a monopsonistic labour market, there are a number of ways in which it is too simplistic, and it is important to be aware of its limitations. First, we have assumed that the employer is a simple monopsonist who must pay the same wage to all workers – that is, wage discrimination is not allowed by assumption.

Second, the simple model assumes that the only way an employer can raise employment is by raising the wage paid, something that is quite implausible. Manning (2006) considers the case where employers can also increase their employment by spending resources on recruitment

activities. He shows that monopsony can be thought of as the case where the marginal cost of recruiting an extra worker is increasing in the number of workers recruited.

Third, the simple model is a model of partial equilibrium – it ignores the interactions with other employers that are very important in reality. One would expect the actions of other employers to affect the labour supply curve facing an individual firm; for example, if other firms pay higher wages we would expect the labour supply to this firm to fall for a given wage. Taking account of these interactions is particularly important when considering the impact of policies like the minimum wage that will affect all employers in a market. Manning (2003a, ch. 12) shows that, while in the simple monopsony model a just-binding minimum wage always raises employment, this is not necessarily the case in general equilibrium models of oligopsony, where there is more than one employer.

The Sources of Monopsony Power

Labour economists have often doubted whether many employers have significant monopsony power over their employees (though this scepticism has diminished in recent years – see Boal and Ransom 1997, for a generally sympathetic survey). So it is important to think about why employers are likely to have monopsony power over their workers.

Traditionally, employers are thought to have monopsony power only in labour markets in which there is a small number of employers. A typical example would be a mill town or mine village in the early days of industrialization, where the employer dominated the local labour market. Most economists are rightly sceptical of the view that the number of employers in many labour markets is small. Classical monopsony could also occur when there are many employers but they collude in wage-setting so that there are only a few effective employers in the labour market. But most economists do not think employer collusion is important in labour markets. (Yet Adam Smith 1776, p. 169, strongly believed that employer collusion was a frequent outcome in labour markets:

> . . . we rarely hear, it has been said, of the combinations of masters, though frequently of those of workmen. But, whoever imagines, upon this account, that masters rarely combine, is as ignorant of the world as of the subject. Masters are always and everywhere in a sort of tacit, but constant and uniform combination, not to raise the wages of labour above their actual rate. To violate this combination is everywhere a most unpopular action, and a sort of reproach to a master among his neighbours and equals. We seldom, indeed hear of this combination, because it is the usual, and one may say, the natural state of things.)

However, modern theories of monopsony do not generally argue that employer market power over their workers derives from there being a small number of employers. They tend to emphasize the role of frictions in the labour market. The perfectly competitive model implies that an employer who cuts wages by one cent will find all their existing workers quit immediately. While it is likely that cutting wages will increase the quit rate and make it harder to recruit replacements, these effects are not as strong as the perfectly competitive model would have us believe.

To illustrate how this can lead to a model from the perspective of firms that looks something like Fig. 1, suppose that the quit rate of workers is a negative function of the wage, $q(w)$ and the flow of recruits to the firm is a positive function of the wage, $R(w)$. Then, in steady state employment in the firm is:

$$N = \frac{R(w)}{q(w)} \tag{4}$$

which will be a positive function of the wage – that is, the employer will face an upward-sloping labour supply curve as represented in Fig. 1.

What are the sources of these frictions in labour markets? In *The Economics of Imperfect Competition* (1933, p. 296), Joan Robinson argued that ignorance (about what all employers are offering), heterogeneous preferences and mobility costs are the most plausible sources of frictions in the labour market. The formal models of recent years are built on these ideas. Models based on worker ignorance are typically search models (the canonical versions

of which are probably Albrecht and Axell 1984, and Burdett and Mortensen 1998) in which it takes time and/or money for workers to change jobs. One the other hand, there are the models that assume workers have full information and no mobility costs but that jobs are differentiated in some way (a canonical model of this sort is Bhaskar and To 1999, though all such models have roots in the model of product differentiation by Salop 1979). In these models, jobs might be differentiated by physical location or skill or any other plausible characteristic. This product differentiation gives employers some monopsony power over their workers because employers are not perfect substitutes from the perspective of workers, so a cut in the wage does not cause all workers to leave for other firms.

These theories of 'modern monopsony' might appear to be very different to classical models of monopsony, but Manning (2003b) argues that they are more similar than one might have thought as they all use different mechanisms to argue that the choice of employers of a particular worker is limited at a particular moment in time.

It is plausible to think that labour markets have frictions; but is this any more than a complication? The next three sections argues that it does matter, emphasizing how our analysis of labour markets from the perspective of workers, employers and public policy is affected in important ways by the recognition that employers have monopsony power over their workers.

Monopsony from an Employer Perspective

Here the key idea of the monopsony model is that the labour supply curve facing an individual employer is not perfectly elastic. It is helpful to think about the decisions employers must make about pay, the structure of pay and non-wage aspects of jobs.

First, monopsonistic employers who want, for whatever reason, to be large will have to pay higher wages as they need to be further up their labour supply curve. Hence monopsony offers a simple explanation for the very robust empirical correlation between employer size and wages (see Brown and Medoff 1989). It can also explain why wages seem to be positively correlated with measures of how 'good' an employer is like productivity and profitability (for example, Blanchflower et al. 1996). As noted in the previous section, 'good' firms that have a higher MRPL curve will choose to pay higher wages, something that should not happen in a perfectly competitive labour market.

We also have robust evidence that low-wage employers find it harder to recruit and retain workers, as predicted by monopsony. Low-wage employers have higher vacancy rates, take longer to fill vacancies and have higher quit rates among their workers.

As already mentioned, employers have an incentive to wage discriminate, to pay different wages to workers who might have the same level of productivity. In particular, we would expect them to pay wages that rise with seniority, since pushing the rewards of employment into the future helps to deter quits as workers get the high wages only if they remain with the firm (see Stevens 2004). This is consistent with the empirical evidence (admittedly a bit patchy) that pay varies more strongly than productivity with seniority, though there are also incentive theories that make similar predictions.

Monopsony also offers a simple explanation of why employers often seem to pay for general training of their workers. In a perfectly competitive market this is something of a puzzle; since workers should receive all the returns to general training, employers should not be prepared to pay for it. But in a monopsonistic labour market, where wages are below marginal products, some of the returns to general training are likely to accrue to employers, giving them an incentive to provide training.

Monopsony from a Worker Perspective

From the perspective of workers, a monopsonistic labour market will appear to be one in which there is heterogeneity in the jobs available (definitely in the wage but quite likely in other dimensions as well) and jobs are hard to find, so getting and

losing jobs are occasions for joy and sadness. If one wants a formal model to capture these ideas, a search model is the right conceptual framework to use. Of course, one can use search models to think about workers' choices whenever they face a distribution of wages even if the origin of that distribution is not the monopsony power of employers, so this area of research is not distinctive to monopsony.

First, it can explain the existence of wage dispersion even in very tightly defined labour markets. This violation of the 'law of one wage' was first documented by the so-called neo-realist labour economists (see Kaufman 1988) in the United States in the 1940s, but most subsequent studies have confirmed it (for example, Groshen 1991). This wage dispersion is exactly what we would expect to see in a monopsonistic labour market in which different employers will choose different wages even if faced with the same labour supply curve. This can then help to explain why high-wage workers are, other things equal, less likely to quit and less likely to be looking for another job as these workers have been lucky enough to find themselves in one of the good jobs in their segment of the labour market.

Second, monopsony can explain part of the rapid growth in earnings over the early stages of the life cycle (as first identified by Mincer 1974). The human capital explanation of this is that workers are accumulating skills but monopsony/search suggests that workers are working themselves into the best jobs in the market (what might be called the accumulation of search capital, the knowledge of which employers pay higher wages). Consistent with this, Topel and Ward (1992) find that one-third of the wage growth of young men in the US labour market is the result of job mobility.

Third, monopsony can explain the earnings losses suffered by displaced workers. It is well-documented that workers who lose their jobs through no fault of their own (for example, through plant closure) tend to suffer losses in earnings (see Kletzer 1998, for a review) and the losses do not completely fit the pattern suggested by human capital theory – in particular, older workers suffer greater losses, even when we control for job tenure.

Monopsony can also explain systematic wage differentials between workers, even if they do not differ in their productivity. For example, if women are less attached to market employment or their decisions on which jobs to take are less motivated by money (Manning 2003a, provides evidence on both these points), then women will earn less than men even if the wage offer distribution they face is the same. The reason is that women will find it harder to accumulate search capital. There may also be incentives for employers to then pay lower wages to women, giving a further twist to their earnings disadvantage. Ransom and Oaxaca (2005) provides some evidence that the quit rate for women is less sensitive to the wage than is the quit rate for men.

Monopsony also has implications for the incentives to acquire human capital. Because the wage is below the marginal product, it is quite likely that some of the returns to investments in human capital accrue to future employers of the worker, though the interests of these employers are not internalized in the education decision. Hence, the social return to education is likely to exceed the private return, leading, in a free market, to underinvestment.

M

Monopsony from a Public Policy Perspective

Thinking of labour markets as pervasively monopsonistic rather than perfectly competitive has implications for how one thinks about the likely effects of interventions in the labour market. In a perfectly competitive labour market one tends to think of the free market outcome as efficient, of any intervention as causing some inefficiency and justifiable only on equity grounds, especially if the equity effect is large and/or the efficiency cost is small. In contrast, if the labour market is monopsonistic, then there is no presumption that the free market is efficient and interventions might be justifiable on efficiency grounds alone. Based on the simple textbook model of a monopsonist presented earlier, one might be tempted to go further and argue that, because wages are below marginal products, interventions to raise wages must, over some

range, improve efficiency. However, in more sophisticated models of monopsony or models of oligopsonistic competition, such a simple conclusion is not necessarily valid. So the monopsonistic approach does suggest approaching the analysis of the impact of interventions with a more open mind than a true believer in perfect competition might be inclined to do.

A good example is the employment impact of the minimum wage. If the labour market is perfectly competitive, one can prove with nothing more than pencil and paper that the minimum wage must reduce employment, and the only purpose of empirical analysis is to decide on how large the reduction is. However, a monopsony approach suggests going to the data with a less certain view about the 'right' answer. The intuition is that, while a rise in the minimum wage reduces the profitability of employing workers for firms, it increases the incentives for workers to work, and the net effect on employment depends on whether the 'demand' or 'supply' effect is dominant. Hence monopsony can explain why the empirical literature often fails to find evidence that it reduces employment (Card and Krueger 1995; Dickens et al. 1999).

Another good example of apparently 'perverse' employment effects can be found in the impact of equal pay legislation. In the UK, this legislation led to a big increase in the pay of women relative to that of men but did not, as the perfectly competitive model would predict, lead to big falls in the relative employment of women (Manning 1996).

Conclusions

There are good reasons to believe that employers have some monopsony power over their workers. Assuming labour markets are monopsonistic also brings the thinking of labour economists in line with the way in which agents perceive the workings of labour markets. Workers do not perceive labour markets as frictionless and changing; getting and losing jobs are routinely reported as major life events. And employers perceive they have discretion over the wages paid, as a reading of any human resource management textbook can confirm. And, as demonstrated in this article, a whole range of puzzles and anomalies melt away once one adopts the monopsony perspective. However, the impact of regulations is more ambiguous than in perfectly competitive markets, and the theoretical perspective should go hand-in-hand with an open-minded empirical approach.

There is much work yet to be done. For example, the size of the wage elasticity of the labour supply curve to an individual firm is very much unknown. The literature on the subject is small and not entirely convincing. The best estimates we do have (probably from Staiger et al. 1999; Falch 2001; Clotfelter et al. 2006) suggest quite a low wage elasticity, with the implication that employers do have significant monopsony power.

See Also

- ▶ Labour Market Search
- ▶ Minimum Wages
- ▶ Robinson, Joan Violet (1903–1983)
- ▶ Wage Inequality, Changes in

Bibliography

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

Bhaskar, V., and T. To. 1999. Minimum wages for Ronald McDonald monopsonies: A theory of monopsonistic competition. *Economic Journal* 109: 190–203.

Blanchflower, D., A. Oswald, and P. Sanfey. 1996. Wages, profits, and rent-sharing. *Quarterly Journal of Economics* 111: 227–251.

Boal, W., and M. Ransom. 1997. Monopsony in the labor market. *Journal of Economic Literature* 35: 86–112.

Brown, C., and J. Medoff. 1989. The employer size–wage effect. *Journal of Political Economy* 97: 1027–1059.

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

Card, D., and A. Krueger. 1995. *Myth and measurement: The new economics of the minimum wage*. Princeton: Princeton University Press.

Clotfelter, C., E. Glennie, H. Ladd, and J. Vigdor. 2006. Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. Working paper no. 12285. Cambridge, MA: NBER.

Dickens, R., S. Machin, and A. Manning. 1999. The effects of minimum wages on employment: Theory and evidence from Britain. *Journal of Labor Economics* 17: 1–22.
Falch, T. 2001. Estimating the elasticity of labor supply utilizing a quasi–natural experiment. Working paper, Norwegian University of Science and Technology, Trondheim.
Groshen, E. 1991. Sources of intra-industry wage dispersion: How much do employers matter? *Quarterly Journal of Economics* 106: 869–884.
Just, R., and W. Chern. 1980. Tomatoes, technology, and oligopsony. *Bell Journal of Economics* 11: 584–602.
Kaufman, B. 1988. *How labor markets work*. Lexington: Lexington Books.
Kletzer, L. 1998. Job displacement. *Journal of Economic Perspectives* 12(1): 115–136.
Manning, A. 1996. The equal pay act as an experiment to test theories of the labour market. *Economica* 63: 191–212.
Manning, A. 2003a. *Monopsony in motion: Imperfect competition in labor markets*. Princeton: Princeton University Press.
Manning, A. 2003b. The real thin theory: Monopsony in modern labour markets. *Labour Economics* 10(2): 105–131.
Manning, A. 2006. A generalised model of monopsony. *Economic Journal* 116: 84–100.
Mincer, J. 1974. *Schooling, experience and earnings*. New York: NBER.
Murray, B. 1995. Measuring oligopsony power with shadow prices: U.S. markets for pulpwood and sawlogs. *Review of Economics and Statistics* 77: 486–498.
Ransom, M., and R. Oaxaca. 2005. Sex differences in pay in a 'new monopsony' model of the labor market. Discussion paper no. 1870. Bonn: IZA.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Salop, S. 1979. A model of the natural rate of unemployment. *American Economic Review* 69: 117–125.
Schroeter, J. 1988. Estimating the degree of market power in the beef packing industry. *Review of Economics and Statistics* 70: 158–162.
Smith, A. 1776. *The wealth of nations*. London: Penguin, 1986.
Staiger, D., J. Spetz, and C. Phibbs. 1999. Is there monopsony in the labor market? Evidence from a natural experiment. Working paper no. 7258. Cambridge, MA: NBER.
Stevens, M. 2004. Wage-tenure contracts in a frictional labour market: Strategies for recruitment and retention. *Review of Economic Studies* 71: 535–551.
Thornton, R. 2004. Retrospectives: How Joan Robinson and B. L. Hallward named monopsony. *Journal of Economic Perspectives* 18(2): 257–261.
Topel, R., and M. Ward. 1992. Job mobility and the careers of young men. *Quarterly Journal of Economics* 107: 439–479.

Monotone Mappings

Peter Newman

In a seminar given in 1934 Abraham Wald was the first to use the weak axiom of revealed preference (1936a, b). However, in several ways his use of it looks odd to the post-Samuelson reader. First, since Wald's main purpose was to establish a new condition for unique solution of the modified Walras–Cassel system of general equilibrium he had introduced earlier (1935), as originally stated it was a restriction on *market* rather than individual behaviour. Secondly, the axiom referred not to the (vector) market demand function $z = f(p)$ but to its inverse $p = f^{-1}(z)$, whose existence is of course quite suspect. Finally, although later in the paper Wald did in fact invoke the individual version (wa) of the weak axiom as some ground – 'a statistical probability' – for belief in its market version (WA), he did not justify (wa) as did Samuelson (1938), as in its own right a sensible rule for consistent market behaviour. Instead, Wald derived if from an assumed additive Jevonian utility function for the individual, i.e. $u_i(z_i) = \Sigma_j u_{ij}(z_{ij})$ where in addition $d^2u_{ij}(z_{ij})/dz_j^2 < 0$ for each person i and good j; and so in Wald (wa) appeared as much more restrictive than it really is.

All this is well known (see e.g. Dorfman et al. 1958, Ch. 13). What is not so commonly remarked is that in the same paper Wald introduced a further assumption (6′) observing it to be stronger than (WA) but weaker than his original assumptions in (1935) to guarantee uniqueness of equilibrium, namely: that each $f_j^{-1}(z_j)$ should exist and be strictly monotone decreasing. The new (6′) was as follows: for any pair of vectors (p, z) with $p = f^{-1}(z)$, let an arbitrary perturbation $\Delta z \neq 0$ occur. Then the resulting vector Δp of price changes, given by $\Delta p = f^{-1}(z + \Delta z) - f^{-1}(z)$, must satisfy

$$\langle \Delta p, \Delta z \rangle < 0 \qquad (1)$$

M

where $< . , . >$ denotes inner product (in the Euclidean space R^n).

Later, in a letter of January 1940 to Karl Menger (partly published in Wald 1952), he showed that (1) holds if the Jacobian of f^{-1} is everywhere negative definite and all the second derivatives of each f_j^{-1} are everywhere continuous. Apart from this, however, he neither then nor later provided any economic rationale for the validity of (1).

The condition (6′) can be reinterpreted if one starts from f rather than f^{-1}. Given (p, z) as before, let $\Delta p \neq 0$ be the independent perturbation and require instead that the resulting Δz [$= f(p + \Delta p) - f(p)$] satisfy (1). Put this way, (1) looks very much like Hicks's 'ultimate generalization of the theory of demand' (1946, pp. 51–2, 329–33; not in the first (1939) edition). Indeed, Baumol and Goldfeld (1968, p. 269) claim that Wald's (6′) fully anticipated Hicks on this important inequality. This seems wrong on three counts. First, Wald's analysis refers to market behaviour and Hicks's to individual behaviour. Secondly, in its original form (6′) is a restriction on f^{-1} and not directly on f. Thirdly and most importantly, Hicks unlike Wald required also that both z and $z + \Delta z$ lie on the same indifference surface, so that (1) is then a restriction not on the ordinary (Marshallian) demand function f but on the compensated (Hicksian) demand function h. In this case (1) follows from the assumption that the individual chooses a unique bundle z to minimize expenditure for any prescribed level of utility.

For as Samuelson (1946–1947) pointed out long ago, if z^1 is known to minimize a linear functional $\langle p, \cdot \rangle$ over some set S^0 to which it belongs (e.g. the 'better set' $B^0 = \{z: z \gtrsim z^0\}$ for some 'target' bundle z^0), while $z^2 \in S^0$ similarly minimizes $\langle p^2, \cdot \rangle$ over S^0, then simply by definition of a minimizing point,

$$\langle p^1, z^1 - z^2 \rangle \leq 0 \text{ and } \langle p^2, z^1 - z^2 \rangle \geq 0 \quad (2)$$

Writing Δp for the linear functional $p^1 - p^2$ and putting $\Delta z \equiv z^1 - z^2$, (1) implies

$$\langle \Delta p, \Delta z \rangle \leq 0 \quad (3)$$

Then (1), which is the strict version of (3), follows from the additional assumption that z^1 or (non-exclusively) z^2 is a unique minimizer of its respective linear functional over S^0.

Is (3) valid in circumstances other than the problem of minimization just described? To ask this is to ask if f, or equivalently f^{-1}, is a *monotone operator*.

Some Definitions

In what follows, X and Y are paired topological vector spaces; the case $X = R^n = Y$ is probably that of chief interest to most readers. A *multifunction* (synonymously, set-valued mapping, correspondence) from X to Y is then a mapping T which for each $x \in X$ assigns a subset $T(x) \subset Y$. In effect, it is an ordinary function from X to the space 2^Y, consisting of all subsets of Y.

The *graph* of T is a set $G(T) \subset X \times Y$, defined by

$$G(T) = \{(x, y) : (x, y) \in X \times Y, y \in T(x)\} \quad (4)$$

The graph $G(T^{-1})$ of the mapping T^{-1}: $Y \to z^X$ which is inverse to T is clearly the same as $G(T)$. Given any family $\{T_j\}$ of multifunctions from X to Y, a partial ordering $\succ$ of $\{T_i\}$ is obtained by writing $T_i \succ T_j$ iff $G(T_i) \supset G(T_j)$ Thus a mapping T_h is *maximal* in the family $\{T_i\}$ if for no $T_j \in \{T_i\}$ is $G(T_h)$ a strict subset of $G(T_j)$.

If for any two $x^1, x^2 \in X$ and any $y^1 \in T(x^1), y^2 \in T(x^2)$ we have

$$\langle x^1 - x^2, y^1 - y^2 \rangle \geq 0 \quad (5)$$

then T is a *monotone mapping* from X to Y. If $-T$ is monotone, then T is said to be *dissipative*. If $x^1 \neq x^2$ implies strict inequality in (5), then T is *strictly monotone*. If T is actually a function from X to Y, i.e. $T(x)$ is a singleton for each x, then T is a monotone *operator*. The theory of monotone mappings and operators is an important though relatively recent branch of nonlinear functional analysis and already has a vast literature, whose terminology has not yet been standardized; see e.g. Browder (1976), Deimling (1985, Chs 3 and 8), Dolezal (1979), Ghizzetti (1969), Joshi and

Bose (1985, Ch. 3), Vainberg (1973) and the references contained therein.

Suppose now that T has the further property that for any finite set of pairs $(x_i, y_i) \in G(T)$ where $i = 0, 1, 2, \ldots, m$ (m arbitrary),

$$\begin{aligned}&\langle x^1 - x^0, x^0\rangle + \langle x^2 - x^1, y^1\rangle \\ &+\ldots \qquad\qquad \ldots + \langle x^m - x^{m-1}, y^{m-1}\rangle \\ &+\langle x^0 - x^m, y^m\rangle \le 0\end{aligned} \tag{6}$$

Then T is said to be *cyclically monotone*, a concept due to Rockafellar (1966); *maximal* cyclically monotone mappings are defined analogously to maximal monotone mappings.

It is obvious from (6) and (5) that if T is cyclically monotone it is monotone, but the reverse is false unless $X = R = Y$ (Rockafellar 1970a, p. 240).

Some Examples

(a) If $f: R \to R$ is monotone increasing, then it is a monotone operator.

(b) Let $f: R^n \to R^n$ be given by $y = Ax$, where A is a (square) matrix, so that f is linear. Then f is a monotone (or dissipative) operator iff A is positive (or negative) semidefinite.

(c) If $X = R^n = Y$ and T_0 is that multifunction associated with a 'better set' B^0 such that for any $z^k \in B^0$, $T_0(z^k) = \{-p^k: z^k$ minimizes $\langle p^k, \cdot\rangle$ over $B^0\}$, then it is easy to check that T_0 is cyclically monotone.

(d) Let X be a real Banach space and Y be its topological dual (i.e. $Y = X^*$). Then $T: X \to 2^X$ is a maximal cyclically monotone mapping iff it is the subdifferential ∂f (see DUALITY) of some proper lower semicontinuous (lsc) function $f: X \to [-\infty, \infty]$ In this case, T determines f up to an additive constant (Rockafellar 1966).

(e) Let X and Y be as in (d), and $f: X \to [-\infty, \infty]$ be convex, lsc and proper. Then its subdifferential ∂f is a *maximal monotone* mapping (Rockafellar 1970b). (Since there are more monotone mappings than cyclically monotone mappings this result is not included in (d), unless $X = R = Y$.)

Monotone Mappings and Revealed Preference

First, it will be shown that if the demand function is a monotone operator then Samuelson's Weak Axiom holds (a result first proved in Unger 1974). Let an individual have income $\mu > 0$ face market prices $p \in R^{n+}$ have a demand function f which is homogeneous of degree zero in prices and income, and satisfy the budget equation strictly. Define new normalized and (personalized!) prices $q = (\mu^{-1})p$. Then a new demand function g can be defined on a suitable convex subset of these prices by $g(-q) = f(p, \mu)$.

As in Samuelson (1938), a bundle z^1 is *revealed preferred* ρ to another bundle z, if $z^1 = g(-g^1)$ and $1 = \langle q^1, z^1\rangle \ge \langle q, z\rangle$. Then the (non-strict) version of (wa) is

$$\begin{aligned}&\left[z^1 p z^2 \text{ and } z^2 = g\left(-q^2\right)\right] \text{ implies} \\ &\qquad\left[\left\langle q^2, z^1\right\rangle \ge \left\langle q^2, z^2\right\rangle = 1\right]\end{aligned} \tag{7}$$

The proof will be by contraposition; (7) is false iff

$$\left\langle q^1, z^1\right\rangle \ge \left\langle q^1, z^2\right\rangle \text{ and } \left\langle q^2, z^2\right\rangle > \left\langle q^2, z^1\right\rangle$$

which may be written

$$\begin{aligned}\left\langle -q^1, z^1\right\rangle \le \left\langle -q^1, z^2\right\rangle \text{ and } &\left\langle -q^2, z^2\right\rangle \\ &< \left\langle -q^2, z^1\right\rangle\end{aligned}$$

On addition these inequalities yield

$$\left\langle -q^1, z^1 - z^2\right\rangle + \left\langle -q^2, z^1 - z^2\right\rangle < 0$$

or

$$\left\langle \left(-q^1\right) - \left(-q^2\right), z^1 - z^2\right\rangle < 0 \tag{8}$$

This contradicts (5), so g is not a monotone operator and the result is proved.

A counter-example that borrows some demand functions from Hurwicz and Richter (1971, p. 65) will show now that the reverse proposition is false. Let there be two commodities z_1 and z_2, with demand functions $f_1(p_1, p_2, \mu) = (p_1/p_2) +$

1 and $f_2(p_1, p_2, \mu) = \left(p_2^{-1}\right)\left(\mu - \left(p_1^2/p_2\right) - p_1\right)$ Consider the one-parameter price-income path beginning at $\left(p_1^1, p_2^1, \mu\right) = (1, 1, 7/3)$ and ending at $\left(p_1^2, p_2^2, \mu\right) = (1, 2, 7/3)$ given by

$$P_j(t) = p_j^1 + t\left(p_j^2 - p_j^1\right) \quad t \in [0,1], j = 1,2 \tag{9}$$

Along this path each price vector p and each bundle z can be regarded as a function of t alone. It is easily shown that $t^r < t^s$ implies $z(t^r)$ ρz (t^s) and that (wa) holds everywhere along the path, as indeed does the Strong Axiom (sa) of Houthakker (1950). Finally, along the path the normalized prices q are simply $q_1 = 3/7$ and $q_2 = (3/7)(1 + t)$.

Let X and Y be a Banach space and its dual respectively, and F a function from a convex set $K \subset X$ to Y Then Kachurovskii (1968, Theorem 1a) showed that F is a monotone operator iff for any $x \in K$ and any $v \in X$ such that $(x + v) \in K$ the function ϕ:[0,1] $\rightarrow R$ is non-decreasing, where ϕ is defined by

$$\phi(t : x, v) = \langle F(x + tv), v \rangle \tag{10}$$

To apply this result to our counter-example, simply put $F = g, -g(0) = x$, and $[(-q(1))$- $- (-q(0))] = v$. Then, using the assumed values,

$$\phi[t : -q(0), q(0) - q(1)] = \langle [z_1(t), z_2(t)], (0, -3/7) \rangle \tag{11}$$

Calculation shows that ϕ attains a unique global minimum on [0, 1] at $t = 1/2$. Indeed, $\phi(0) = -0.1429$, $\phi(1/2) = -0.1905$, and $\phi(1) = -0.1786$.

Hence, by Kachurovskii's theorem g is not a monotone operator, even though (wa) holds throughout. Since cyclic monotonicity implies monotonicity, g is not cyclically monotone either. So this example does double-duty, showing also that (sa) does not imply cyclic monotonicity. Thus the statements by Jorgenson and Lau (1974, p. 190), asserting the equivalence of monotonicity and (wa), cyclic monotonicity and (sa), are not correct.

Finally, consider demand functions where income is replaced by initial endowments z^0 valued at current prices, i.e. $z = \Psi\left(p, \langle p, z^0 \rangle\right)$ A counter-example very similar to the last one may be used to show that (wa) does not imply that Ψ is a monotone operator, contradicting a claim of Kusumoto (1977, p. 1942).

Conclusion

It is remarkable that Wald formulated the monotonicity requirement (6′), long before monotone mappings became an established branch of nonlinear functional analysis with the work of Minty (1962) and others. Given that propitious early start, it would be fitting if economists could go on to make use of the many powerful results in the theory of monotone mappings that have been obtained since Wald's time. Although the subdifferentials of convex functions form one very important subclass of such mappings prominent in economic analysis, the additional results for these from monotonicity theory *per se* seem to add little to the results already obtainable from convexity theory. Whether there are *other* significant monotone functions and correspondences in economics remains an open question.

See Also

- ▶ Convex Programming
- ▶ Correspondences
- ▶ Duality
- ▶ Integrability of Demand
- ▶ Revealed Preference Theory
- ▶ Uniqueness of Equilibrium

Bibliography

Baumol, W.J., and S.M. Goldfeld (eds.). 1968. *Precursors in mathematical economics: An anthology*, Series of reprints of scarce works in political economy, vol. 19. London: London School of Economics and Political Science.

Browder, F.E. 1976. *Nonlinear operators and nonlinear equations of evolution in Banach Spaces*. Proceedings

of Symposia in Pure Mathematics, Vol. 18, Part 2. Providence: American Mathematical Society.
Deimling, K. 1985. *Nonlinear functional analysis*. Berlin: Springer.
Dolezal, V. 1979. *Monotone operators and applications in control and network theory.* Amsterdam: Elsevier Scientific.
Dorfman, R., P. Samuelson, and R. Solow. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill.
Ghizzetti, A. (ed.). 1969. *Theory and application of monotone operators*. Gubbio: Edizioni Oderisi.
Hicks, J.R. 1946. *Value and capital*, 2nd ed. Oxford: Clarendon Press.
Houthakker, H.S. 1950. Revealed preference and the utility function. *Economica* NS 17: 159–174.
Hurwicz, L., and M.K. Richter. 1971. Revealed preference without demand continuity assumptions. Ch. 3. In *Preferences, utility and demand*, ed. J. Chipman, L. Hurwicz, M.K. Richter, and H. Sonnenschein. New York: Harcourt Brace, Jovanovich.
Jorgenson, D.W., and L.J. Lau. 1974. The duality of technology and economic behavior. *Review of Economic Studies* 41: 181–200.
Joshi, M.C., and R.K. Bose. 1985. *Some topics in nonlinear functional analysis*. New York: Wiley.
Kachurovskii, R.I. 1968. Non-linear monotone operators in Banach spaces. *Russian Mathematical Surveys* 23: 117–165.
Kusumoto, S.I. 1977. Global characterization of the weak Le Chatelier-Samuelson principle and its applications to economic behavior, preferences and utility-'embedding' theorems. *Econometrica* 45: 1925–1956.
Minty, G.J. 1962. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal* 29: 341–346.
Rockafellar, R.T. 1966. Characterization of subdifferentials of convex functions. *Pacific Journal of Mathematics* 17: 497–510.
Rockafellar, R.T. 1970a. *Convex analysis*. Princeton: Princeton University Press.
Rockafellar, R.T. 1970b. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics* 33: 209–216.
Samuelson, P.A. 1938. A note on the pure theory of consumer's behaviour. *Economica* NS 5: 61–71.
Samuelson, P.A. 1946–47. Comparative statics and the logic of economic maximizing. *Review of Economic Studies* 14: 41–43.
Unger, K. 1974. Monotone operators and revealed preference. PhD dissertation, Johns Hopkins University.
Vainberg, M.M. 1973. *Variational method of monotone operators in the theory of nonlinear equations*. New York: Wiley.
Wald, A. 1935. Über die eindeutige positive Losbarkeit der neuen Produktionsgleichungen (I. Mitteilung). *Ergebnisse eines Mathematischen Kolloquiums, 1933–4*. 6: 12–18.
Wald, A. 1936a. Über die Produktionsgleichungen der ökonomischen Wertlehre (II. Mitteilung). *Ergebnisse eines Mathematischen Kolloquiums, 1934–5* 7: 1–6. Wald (1935, 1936) are translated as chapters 25 and 26, respectively, in Baumol and Goldfeld (1968).
Wald, A. 1936b. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7: 637–670. Trans. by O. Eckstein as Some systems of equations in mathematical economics. *Econometrica* 19, October 1951:368–403.
Wald, A. 1952. On a relation between changes in demand and price changes. *Econometrica* 20: 304–305.

Montchrétien, Antoyne de (1575–1621)

P. Bridel

Montchrétien's place in the history of economics is probably more the result of the title than the content of his 1616 *Traicté de l'oeconomie politique* – never before had the words 'political' and 'economy' been put together on the title page of a volume claiming to be a treatise, that is, dealing systematically with one subject. For some, this is Montchrétien's only merit; for others, painstakingly sorting the analytical wheat from the factual chaff. Montchrétien's contribution to economics, if somewhat lacking in originality, brings forward for the first time some important elements of what was to become standard Mercantilist thinking.

Sharing the political credo of his contemporary Jean Bodin, Montchrétien is, however, the first to add (to foreign wars) the *search for wealth* as a means to keep stable France's social order organized around the king. The *Traicté* is among the first works to question explicitly the old Aristotelian argument about the independence of politics from (and its superiority over) all other aspects of social life, including economic activities.

Labour being no longer a curse, but one of the pillars of political stability, productive labour and the search for wealth are the logical conclusions to which Montchrétien is led: 'Le bonheur des hommes . . . consiste principalement en la richesse, et la richesse dans le travail' (1616, p. 99).

Besides agriculture, industry and trade are given the pride of place by Montchrétien in his analysis of the workings of the 'social body'. Since exchange is the ultimate *raison d'être* of most productive labour, traders and 'marchands' play a central coordinating role. Profit, being their incentive, has to be encouraged and protected:

> . . . les marchands sont plus qu'utiles . . . et . . . leur souci de profit, qui s'exerce dans le travail et l'industrie, fait et cause une bonne part du bien public . . . Que, pour cette raison on leur doive permettre l'amour et la quête du profit . . . (1616, p. 137).

The Mercantilist argument about the necessity for governments to help increase the wealth of nations follows naturally. Having thus underlined for the first time the close relationship between politics and economics, it remained for Montchrétien to christen *political economy* this body of primitive arguments that was to study systematically, a century and a half later, how the wealth of nations is produced, distributed and exchanged.

Selected Works

1616. *Traicté de l'oeconomie politique*, ed. Th. Funck-Brentano. Paris: Plon, 1889.

Monte Carlo Methods

John G. Cragg

The term 'Monte Carlo methods' is used to refer to two different, though closely related, techniques. The first meaning, currently the less common one among economists, is the evaluation of definite integrals by use of random variables. The idea is to evaluate $\int_a^b F(x)\mathrm{d}x$ (where x may be a vector) by estimating $\int_a^b [F(x)p(x)]p(x)\mathrm{d}x$. Here $p(x)$ is the density function of a random variable defined over $[a, b]$. The original problem has been converted into one of estimating the mean of $F(x)/p(x)$. It can be solved by using a random sample drawn from $p(x)$ and calculating the average value of $F(x)/p(x)$.

Despite widespread occurrence of intractable integrals in economics, Monte Carlo methods in this sense have been little used (except to the extent that the second meaning can be encompassed within the first), possibly because explicit parameters for $F(x)$ are usually unknown and so only general analytical solutions are of interest. Promising opportunities for application may arise in Bayesian econometrics where fully parameterized (and often very intractable) definite integrals are the rule rather than the exception. A good example of use of Monte Carlo methods in this sense is provided in Kloek and van Dijk (1978).

The second meaning of 'Monte Carlo methods' refers to repeated simulation of a stochastic model to investigate the properties of statistical techniques applied to it. The techniques under investigation typically are derived from general principles such as maximum likelihood which provide no guarantee of reliability in finite samples. The Monte Carlo procedure is adopted when analytical derivation of finite sample properties of methods appears not to be feasible.

More explicitly, suppose that some observable economic variables of interest, y, are generated according to where $y = g\ (\varepsilon;\ x,\theta)$ e ε are unobserved random variables of specified distribution, x are other observable variables on which the analysis of y can be conditioned and θ are unknown parameters. Monte Carlo methods investigate the properties of the statistical techniques which are used to infer properties of the process generating y. They do this by applying the techniques to artificially generated data coming from this model. Repeated samples of ε are obtained from the specified distribution and samples of y are generated using chosen values of x and θ.The techniques being investigated are applied to these artificial sets of data on y to provide samples of inferences made by the techniques. Properties of the procedures may then be established by statistical inference. Since all aspects of the data generation are known, the

extent to which the techniques are reliable can be assessed.

The Monte Carlo method involves substitution of computer resources for human resources of the sort needed to perform abstract mathematical derivations (as the title of the classic paper by Summers (1965) indicated). Despite the increase in technical expertise among economists, progress in computing technology continues to be so rapid that one may well expect application of the technique to become more rather than less common. The volume of work using Monte Carlo methods is already vast and is steadily growing. A good recent survey of some of this literature is Hendry (1984).

Monte Carlo methodology is highly dependent on the ability of computers to generate pseudo-random numbers that mimic the random processes hypothesized to generate economic data. None of the variety of techniques available for this purpose can produce truly independent sets of random numbers. (Kennedy and Gentle 1980, provides a good discussion of random number generation.) As a result there may be a legitimate worry that the lack of independence interacts with the complicated processes in many econometric specifications to produce misleading results. In addition to careful examination of the random number generators used and their properties, one practical way to lessen this danger is to use within a Monte Carlo study different random number generators that employ substantially different methods. The results of the sub-experiments using the different generators can then be compared to establish that they are in agreement with each other. To date such validation seems rare. Indeed, explicit reference to the type of random-number generator being used is not common, despite the known weaknesses of some generators for which efficient computer code has been readily available.

The usual criticisms of the Monte Carlo technique concern the lack of precision of the findings and their dependence on a particular specification. The first is largely a problem of sample size and efficiency of experiment design. A variety of techniques, largely stemming from Hammersley and Handscombe (1964), is available to increase efficiency, of which the most common and generally useful is employment of control variate estimators whose distributions are known and which are likely to be closely related to the techniques of interest. These variance-reducing techniques make use of the fact that all parameters of the generating process are known by the investigator; this knowledge can therefore be exploited in discovering the properties of techniques used when such information is not available. Some authors indeed would restrict the term 'Monte Carlo' to studies that exploit these possibilities to obtain efficiency. However, with the ready availability of cheap computation facilities, high degrees of precision may often be achieved even without such techniques. In so far as more sophisticated experimental design requires more complicated computer programming, the gain in efficiency may be illusory.

The problem of results being dependent on specific parameters is apparently more serious, but can be over-emphasized. Two approaches to reducing the problem can be adopted. First, the parameters can be varied and the effect of the variation can be studied. Although past studies have tended to use many replications at each of a very few points in the relevant parameter space, it may instead be sensible to allow the parameters, θ and/or x, to vary from replication to replication.

With this approach, the problem becomes one of fitting a 'response surface' to describe the ways in which the properties of the econometric techniques depend on the various parameters or conditions of the different replications of an experiment. In principle, it should be possible to discover the properties of the finite-sample distribution to any desired degree of accuracy using standard statistical approximation techniques. Though this might seem to indicate that Monte Carlo techniques can replace exact derivation of the sampling distribution of estimators, the lack of precision and completeness of any Monte Carlo study and the difficulty of finding a revealing and parsimonious response-surface representation in the absence of knowledge of what aspects of the experimental situation are critical should not be minimized. Furthermore, in the past Monte Carlo studies have run into serious problems from not appreciating features of the small sample

distributions of techniques being studied, such as the lack of moments for some simultaneous equation estimators or the difficulties encountered on the unit circle in moving average models. However, exact results, even when available, may also be difficult to interpret or apply.

A second approach to the problem of the specificity of Monte Carlo results is to conduct experiments using parameters fitted to a sample of data that are believed to arise from the process of interest. Thus the values of θ estimated from actual data on y and the corresponding values of x are employed. These provide a presumption that the Monte Carlo experiment is investigating the relevant part of the parameter space. This approach largely overcomes the specificity problem in the sense that the Monte Carlo study can answer the question whether the inferences drawn about the processes generating the data would tend to be made if data had in fact been generated by the supposed process with the estimated parameters. While the results can not be generalized to other applications of the techniques readily, this may not be important if indeed the application is important and interesting and the Monte Carlo investigation is inexpensive.

A major weakness that standard Monte Carlo methodology shares with traditional exact sample results is the specification of particular distributions for ε.It is often doubtful that these describe adequately the process generating economic data. To some extent, 'bootstrap' techniques (cf. Efron 1982), which involve using the empirical distribution of residuals in a study, overcome this, providing a more concrete incorporation of Monte Carlo techniques into the process of making statistical inferences about processes generating data.

A side benefit that Monte Carlo studies may provide is validation of computer software. They may also reveal computational problems of particular methods that are not immediately obvious. Because Monte Carlo studies involve repeated use of estimating or testing techniques, they may uncover programming bugs or computational difficulties which would not surface in only a few applications. Furthermore, results vastly at difference with those expected, for example from asymptotic theory, may in particular instances indicate failures in computer programming rather than weakness of the econometric methods.

References

Efron, B. 1982. *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

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

Hendry, D.F. 1984. Monte Carlo experimentation in econometrics. In *Handbook of econometrics*, vol. II, ed. M. Intriligator, 937–976. Amsterdam: North-Holland.

Kennedy Jr., W.J., and J.E. Gentle. 1980. *Statistical computing*. New York: Marcel Dekker.

Kloek, T., and H.K. van Dijk. 1978. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica* 46: 1–19.

Summers, R.M. 1965. A capital-intensive approach to the small sample properties of various simultaneous equation estimators. *Econometrica* 33: 1–41.

Moore, Henry Ludwell (1869–1958)

A. W. Coats

Keywords

Agricultural economics; Moore, H. L.; Schultz, H.; Statistics and economics

JEL Classifications

B31

An outstanding pioneer econometrician, Moore was a retiring, highly sensitive, intensely dedicated man, who devoted his whole life to the construction of 'a statistical complement to economics', as he termed it. He was born at Moore's Rest, Maryland, on 21 November 1869. After graduating from Randolph Macon College in 1892, he studied under Carl Menger in Vienna, and Simon Newcomb and John Bates Clark at Johns Hopkins, where in 1896 he completed his Ph.D. dissertation on von Thünen's theory of the natural wage. Following a year's instructorship at

Hopkins, and five years at Smith College, Moore taught at Columbia, mainly mathematical economics and statistics, from 1902 to 1929. Essentially a researcher rather than a pedagogue, he attended Karl Pearson's courses on mathematical statistics and correlation in London, in 1909 and 1913, and for several years took a voluntary salary reduction in order to avoid undergraduate teaching. Ill health forced his early retirement.

In a series of powerful and highly original volumes Moore endeavoured, among other things, to verify the marginal productivity of wages, render the Walrasian system statistically operational, and reveal the fundamental law and cause of cycles – wherein he concluded that 'the law of the cycles of rainfall is the law of the cycles of the crops and the law of Economic Cycles' (1914, p. 135). Needless to say, this immensely ambitious undertaking was often severely attacked by contemporaries and subsequent commentators who exposed the data deficiencies, lax hypotheses, unavoidably heroic oversimplifications, and other shortcomings (cf. Stigler 1965, 1968). Nevertheless, the strength and purity of Moore's scientific vision, and the careful and sophisticated statistical methods he employed, commanded respect and admiration.

Not surprisingly, Moore founded no school. Yet his principal disciple, Henry Schultz, was only one among the many economists who produced the 20th-century 'avalanche of statistical demand curves' (Schumpeter 1954, p. 213) inspired by Moore, whose researches exerted a major impact on agricultural economics. Thus Moore may be credited in part with the high scientific standing American agricultural economics now enjoys (Leontief 1971). However, despite his seminal efforts to develop empirical estimates of theoretical economic relationships, Moore's achievements have been insufficiently acknowledged, partly, no doubt, because he was unwilling to propagandize his methods among his fellow professionals.

Selected Works

1908. The statistical complement of pure economics. *Quarterly Journal of Economics* 23: 1–33.
1911. *Laws of wages: An essay in statistical economics*: New York: Macmillan Co.
1914. *Economic cycles: Their law and cause*. New York: Macmillan. (Japanese translation, Tokyo, 1926).
1917. *Forecasting the yield and price of cotton*. New York: Macmillan.
1923. *Generating economic cycles*. New York: Macmillan.
1929. *Synthetic economics*. New York: Macmillan.

Bibliography

Leontief, W. 1971. Theoretical assumptions and nonobserved facts. *American Economic Review* 61: 1–7.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Stigler, G.J. 1965. Henry L. Moore and statistical analysis. In *Essays in the history of economics*, ed. G.J. Stigler. Chicago: University of Chicago Press.
Stigler, G.J. 1968. Moore, H.L. In *International encyclopedia of the social sciences*, ed. L. Sills David, vol. 10. New York: Macmillan and Free Press.

M

Moral Hazard

Y. Kotowitz

Keywords

Adverse selection; Agency theory; Assignment problems; Asymmetric information; Contingent contracts; Contract enforcement; Exclusive contracts; Externalities; Hidden actions; Incomplete contracts; Incomplete information; Insurance; Joint production; Law of large numbers; Licensing; Monitoring; Moral hazard; Noise; Non-cooperative game theory; Non-market institutions; Opportunistic behaviour; Optimal contracts; Rationing; Reputation; Risk aversion; Risk sharing; Spot markets; Transaction price; Vertical integration

JEL Classifications

D8

The problem of moral hazard is pervasive in economic activities. Economists have been well aware of its existence as the following quote from the *Wealth of Nations* will testify:

> The directors of such companies, however, being the managers rather of other peoples' money than of their own, it cannot well be expected, that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own ... Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company. (Smith 1776, p. 700)

However, theoretical developments and their application to specific problems have only proceeded since the 1960s and are still the subject of vigorous research. While we have a considerable understanding of the problem, we do not as yet understand fully market and social responses to it. In the following I shall attempt to explain the nature of the problem and selectively illustrate the flavour of current theoretical developments.

Moral hazard may be defined as actions of economic agents in maximizing their own utility to the detriment of others, in situations where they do not bear the full consequences or, equivalently, do not enjoy the full benefits of their actions *due to uncertainty and incomplete information or restricted contracts* which prevent the assignment of *full* damages (benefits) to the agent responsible. It is immediately apparent that this definition includes a wide variety of externalities, and thus may lead to nonexistence of equilibria or to inefficiencies of equilibria when they exist.

It is a special form of incompleteness of contracts which creates the conflict between the agent's utility and that of others. Such incompleteness may arise due to several reasons: the coexistence of unequal information and risk aversion or joint production, costs and legal barriers to contracting and costs of contract enforcement. We shall analyse each in turn.

Unequal Information

Agents may possess exclusive information. Arrow (1985) classifies such informational advantages as 'hidden action' and 'hidden information'. The first involves actions which cannot be accurately observed or inferred by others. It is therefore impossible to condition contracts on these actions. The second involves states of nature about which the agent has some, possibly incomplete information, information which determines the appropriateness of the agent's actions, but which are imperfectly observable by others. Thus, even if agents' actions are costlessly observable by others, they do not know with certainty whether the actions were in their interest.

Commonly analysed examples of hidden actions are: workers' effort, which cannot be costlessly monitored by employers, precautions taken by insured to reduce the probability of accidents or damages due to them, which cannot be costlessly monitored by insurers. Criminal activity clearly belongs in this category as well.

Examples of hidden information are expert services – such as physicians, lawyers, repairmen, managers and politicians.

Where consequences of specific agents' actions can be separated from those of others, even though the consequences may be affected by random, unobservable states of nature, the problem may be easily solved if agents are risk neutral, by simply assigning the full consequences to the agent, in exchange for a fixed fee. This is in effect a complete contract. The problem of contract incompleteness arises when agents are risk averse or where assignment of responsibility to one agent cannot be made.

When agents are risk averse, assigning full damages (benefits) to them assigns them all risk due to random states of nature. Risk-averse agents would like to purchase insurance against such risks. However, it is impossible for others to separate the consequences of agents' actions from random elements which cannot be controlled by the agent. Insurance against the latter will inevitably insulate agents from the consequences of their own actions. The agent may, of course, offer to supply information about the unobserved actions or states – but such information cannot be credible.

Optimal contracts generally involve some degree of insurance and hence lead to a conflict between incentives and risk sharing. Most of the

literature on moral hazard has concentrated on this case. We shall come back to it.

When precise assignment of responsibility to individual agents is impossible, full assignment of consequences to individual agents cannot be achieved. By definition, this is the case for crime, where the identity of the perpetrator is generally not known with certainty. The design of punishments and the interaction with enforcement activities to apprehend and convict criminals is treated extensively in the literature (see for example Becker 1968).

Group production is another area where assignment may be impossible. Some forms of collective punishment on the group as a whole when output falls short of a specified quota, with some allocation rule when output meets or exceeds the quota may serve to elicit the desired output (Holmstrom 1982). However, the conditions under which this is possible are quite restricted.

Similar problems arise where quality of products is difficult to ascertain because they must be used jointly with another service or product, because their performance is affected by conditions and nature of use. For example drugs must be used in conjunction with physicians' services. Failure of the drug may result from its poor quality, from misdiagnosis by the physician (who may prescribe the wrong drug) or from failure to follow instructions by the patient. In the absence of these complications, it would be optimal for the manufacturer, who knows the quality of his product, to supply a guarantee of performance, in order to remove the incentive to supply lower quality. As well, the guarantee serves at least partly to insure risk averse consumers against random variations in the performance of the drug. Even if the manufacturer is risk-averse, his risk is mitigated by the 'law of large numbers', so it is optimal for him to act as insurer.

However, under the circumstances above such insurance creates a moral hazard problem for the physician and the patient, who may use insufficient care in diagnosis and use. Any risk sharing among the relevant parties therefore induces a moral hazard problem which cannot be avoided in the presence of private information, even if all parties are risk neutral.

Barriers to Contracting

Incomplete contracts may also arise in the absence of private information due to costs of writing detailed contingent contracts. This problem is particularly severe in contracts involving complex transactions and long periods. When uncertainty about the future is great, the number and nature of eventualities to be considered is clearly very large. The cost of anticipating them and writing a contract which specifies or elicits desired actions may be very large. The cost of reaching agreement on the proper actions in each eventuality may well be prohibitive. If the probability of any event is small, and the cost of agreement high, it may pay to leave the contract vague and wait for the resolution of uncertainty before reaching agreement. Of course, this is precisely the case in spot market transactions. However, frequently decisions must be made prior to the resolution of uncertainty. For example, specialized investments in physical or human capital must be made by the parties before production and trading begin (Becker 1964). The nature of the investment may well depend on the transaction price, which may in turn depend on information revealed after the investment is made. A limited agreement on investment and trading may be optimal, leaving transaction price to future negotiation. This, however, may lead to a moral hazard problem. Opportunistic behaviour in subsequent periods by one of the parties may lead to termination of trading or unfavourable contract terms, for the party which invested in specialized capital. Knowing that this may occur, the incentive to invest is reduced. The resulting inefficiency may well fall short of the costs of complete contracts. Williamson (1985) argues that such problems may give rise to vertical integration.

Contracts are too costly to write when transactions are infrequent and small. Most spot market transactions between retailers and consumers falls in this category. Blanket contracts offered by sellers in the form of 'money back guarantees' or exchange privileges may be substituted for explicit contingent contracts – but they are subject to moral hazard on the consumers' side. Alternatively the state legislates fair trading laws which serve as generalized contracts.

Contracts are lacking altogether when transactions are random or involuntary. Accidental damages inflicted on one party by another as in a traffic accident are good examples. Here again, the law must form a generalized contract. It is obvious that such a law cannot possibly allow for all contingencies, so that it constitutes an incomplete contract, giving rise to moral hazard problems. The question of the design of liability rules has been extensively analysed in the law and economics literature (Posner 1977).

Finally, contracts may be restricted by law or by limited financial resources of agents. For example, even if managers are risk neutral, their financial resources may be insufficient to become sole proprietors, without relying on outside capital. Shareholders and bondholders must then share in the risk – raising a moral hazard problem due to the informational advantages of managers. For an extensive analysis of these problems, see Jensen and Meckling (1976).

Similarly, when punishments are limited by law, moral hazard may not be resolved even where actions can be costlessly observed *ex post.* Thus, for example, bankruptcy and limited liability provisions insure borrowers against extremely unfavourable states of nature without limiting the gains from extremely favourable ones. This creates a moral hazard problem, inducing borrowers to undertake riskier projects. Stiglitz and Weiss (1981) show that lenders will sometimes require collateral and ration loans in attempting to overcome these difficulties.

Problems of Enforcement

A related barrier to complete contracting arises from costs and other limitation on enforcement. When enforcement is costly, it may be more efficient to live with the inefficiencies generated by the moral hazard, than to try to enforce the optimal contingent contract. A common way to overcome such difficulties is by way of posting a bond, which is forfeit in the event of non-performance. However, restricted financial resources generally prevent bonding.

Under conditions where enforcement is not economical, contracts must be *self- enforcing.* It is unimportant whether contracts are explicit or implicit, as they frequently are in labour markets. To be viable contracts must make subsequent actions by contracting parties consistent with their self-interest, that is, they must allow for the potential exercise of moral hazard. This problem is at the heart of noncooperative game theory, which defines moral hazard as opportunistic behaviour.

So far we have surveyed the conditions under which a moral hazard problem cannot be trivially resolved. This raises three questions which theorists have begun tackling in the past two decades: (*a*) the nature of optimal contracts in the presence of moral hazard; (*b*) market and institutional/legal response to mitigate these problems; and (*c*) welfare consequences.

Optimal Contracts

The problem has mainly been tackled by agency theory. Following seminal work by Wilson (1969) and Ross (1973) the optimal (typically second best) reward structure for an agent is derived on the basis of observed variables, usually under 'hidden action' assumptions. Some of the main results for risk-averse agents are: (*a*) Optimal contracts require risk sharing between principal and agent which creates a moral hazard problem in the form of insufficient incentives. (*b*) Efficient contracts should utilize all the information available, that is they should be constructed on the basis of statistical inference from the information available on the hidden action of the agent (Holmstrom 1979).Thus monitoring, which reduces inference errors, is productive. (*c*) The nature of the reward schedule is sensitive to the nature of the information available, the residual uncertainty and the degree of risk aversion of the agent and principal. This observation is troubling because incentive contracts observed in reality are generally simple and uniform across a variety of

agents and information sets. Long-term contracts, explicit or implicit (client relations), tend to mitigate moral hazard problems, by introducing a reward for not exploiting short-term informational advantages, and because cumulative information reduces uncertainty. Hence, for example, experience rating in insurance contracts.

Market and Institutional Responses

Market responses may invalidate or reinforce the special features of contracts to mitigate the moral hazard problem. These responses depend on the nature of competition. Free entry and the existence of unobserved differences among agents create the additional problem of adverse selection. We shall therefore reflect only on market responses which are mainly a consequence of moral hazard.

As indicated above, contracts typically require some risk sharing (coinsurance) between the parties when agents are risk averse. Therefore, agents generally bear more risk than they desire. If they are able to purchase additional insurance from third parties, the moral hazard problem is aggravated, making the original contract inefficient. This requires exclusivity in contracting. Thus for example, insurance companies do not allow insurance claims for damage due to fire, health or accident insurance from more than one company. It is obvious that any restriction on coinsurance can be circumvented if such claims are allowed. At the extreme, agents might have more than full coverage, inducing intentional damages, such as arson.

This tendency for exclusivity is reinforced by the advantages of long-term contracting. In the presence of risk aversion or limits on agents' capital which prevent effective bonding, it may be necessary to promise future rewards to mitigate short term opportunistic behaviour. Termination of the agreement will deny these rewards and thus operates as a threat. This requires that contracts yield some rents to agents, so that their removal may constitute a punishment. Thus for example, the utility of being employed must exceed the utility of being unemployed (Shapiro and Stiglitz 1984).

This requires rationing, which is not undone by competition. If being fired by one's employer leads to immediate employment elsewhere at the same wage, rather than to a significant period of unemployment, the threat of firing is ineffective. An equilibrium must be supported by transaction costs of finding new employment or by a collective use of the information contained in the firing. Such information is indeed relevant for hiring decisions by other firms. Its use depends on the costs of obtaining such information. Markets develop to supply such information, thereby increasing the effectiveness of such agreements. Credit information bureaus and employment agencies are some examples. Fama (1980) argues that such 'reputation' mechanisms eliminate moral hazard problems in executive markets. However, as the information is subject to noise, it is clear that moral hazard problems cannot be entirely resolved.

Non-market institutions may develop to mitigate some of these problems. Professional licensing and certification limit the number of physicians, lawyers and many other professionals. Aside from issues of assurance of minimum quality and monopoly, these arrangements insure rents to the professions involved and, hence, make license removal a significant penalty (Arrow 1963).

The consequences of moral hazard in political processes have largely been neglected by economists. Exceptions are Stigler (1971) and Peltzman (1976), who analysed the motivations of regulators, and Buchanan and Tullock (1962). The theoretical tools of agency, contract and game theory have yet to be fruitfully employed in this area. Given the expanding role of government and the evidence of widespread abuses in the political process, such application promises to yield significant dividends.

General Equilibrium and Welfare Effects

There has been little research on the welfare implications of moral hazard. An exception is Stiglitz

(see for example Arnott and Stiglitz 1985), who noted that the existence of moral hazard creates second best contracts. In an economy characterized by such contracts, changes in contracts between any two parties have significant first order effects on social welfare, in contrast to the Arrow–Debreu economy, where first order effects of individual actions are zero at an optimum. As we have seen, moral hazard may lead to rationing and queues, suboptimal expenditure of hidden actions and imperfections in capital markets.

This is not surprising because moral hazard is basically a form of externality. It is well known that uninternalized externalities lead to nonconcavities, possible nonexistence of equilibria and inefficiencies. The existence of such inefficiencies signals a possible role for government. However, government intervention may well cause more problems than it solves. For example, attempts to supplement deficient insurance markets in the form of universal income (social security, income taxation) insurance have run into serious moral hazard problems of work incentives, tax avoidance and evasion, and so on. It is at least partly because of these moral hazard problems that such markets failed to develop. It is therefore unclear whether government supply of these services enhances welfare.

In contrast, government policies which enhance complete contracts and improve their enforcement, can be welfare enhancing. Examples are contract law, liability rules and trade regulations.

See Also

- ▶ Adverse Selection
- ▶ Health Economics
- ▶ Incomplete Contracts
- ▶ Principal and Agent (i)
- ▶ Principal and Agent (ii)

Bibliography

Arnott, R., and J. Stiglitz. 1985. Labor turnover, wage structures, and moral hazard: The inefficiency of competitive markets. *Journal of Labor Economics* 3: 434–462.

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 541–567.

Arrow, K. 1985. The economics of agency. In *Principals and agents: The structure of business*, ed. J. Pratt and R. Zeckhauser. Boston: Harvard Business School Press.

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

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.

Buchanan, J.M., and G. Tullock. 1962. *The calculus of consent*. Ann Arbor: University of Michigan Press.

Fama, E. 1980. Agency problems and theory of the firm. *Journal of Political Economy* 88: 288–307.

Green, J. 1985. Differential information, the market and incentive compatibility. In *Frontiers of economics*, ed. K.J. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

Harris, M., and A. Raviv. 1979. Optimal incentive contracts with imperfect information. *Journal of Economic Theory* 20: 231–259.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74–91.

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13: 314–340.

Jensen, M., and W. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs, and capital structure. *Journal of Financial Economics* 3: 305–360.

Peltzman, S. 1976. Towards a more general theory of regulation. *Journal of Law and Economics* 19: 211–240.

Posner, R. 1977. *The economic analysis of law*. 2nd ed. Boston: Little, Brown.

Ross, S. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63: 134–139.

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

Shavell, S. 1979. Risk sharing and incentives in the principal and agent relationship. *Bell Journal of Economics* 10: 55–73.

Smith, A. 1776. *An inquiry into the nature and the causes of the wealth of nations*. Ed. E. Cannan. New York: Modern Library, 1937.

Stigler, G. 1971. The theory of economic regulation. *Bell Journal of Economics and Management Science* 2: 3–21.

Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71: 393–410.

Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.

Wilson, R. 1969. The structure of incentives for decentralisation under uncertainty. In *La Décision*. Paris: Editions du CNRS.

Moral Philosophy

R. S. Downie

Some idea of the nature of moral philosophy is provided by considering the analogies which philosophers have used over the centuries to explain their aims. This entry will give a brief account of these and then a longer one of the preoccupations of moral philosophers this century.

For Plato (4th century BC), the moral philosopher is the *authoritative guide* to the good life. Asserting the premise that virtue is knowledge, Plato goes on to develop a view on the nature of such knowledge and how it can be acquired. The conclusion is that only certain people are fitted by natural endowment to attain it, and then only after a prolonged period of intellectual, physical and moral education. But, granted they have gone through the process, Plato regards them as fitted to determine the laws and to govern; they are the 'philosopher kings'. Plato is clearly ascribing to moral philosophy the strongest possible normative function, but apart from any philosophical problems attached to his ideas, we have difficulty in a democratic age with his general approach.

Aristotle (4th century BC) seems more plausible. His method is to examine the current views on any topic – what he calls the views of the many and the wise – and then to sift them with the aim of uncovering the general principles embodied in them. Aristotle is therefore inviting us to see the aim of the moral philosopher on the analogy of the *interpreter*. The assumption is that the first principles of conduct are not easily picked out because they are immersed in the details, but the moral philosopher can do so provided he accepts the general opinions of mankind on right and wrong and then uses a process of philosophical analysis to purge these opinions of inconsistency.

A third and totally different analogy is that of the '*razor*'. The principle – *entia non sunt multiplicanda* – commonly associated with the medieval philosopher William of Ockham, has been used as a philosophical razor to cut out what are thought to be redundant concepts, often in practice concepts in ordinary moral and political thinking which cannot be analysed in terms of sense-experience. Concepts such as 'moral obligation' or 'natural law' have sometimes been cut out by those who are using this philosophical razor. There are general difficulties in the basic doctrine, and the artificial limbs which must be manufactured when concepts such as 'moral obligation' are cut off are not more serviceable. Undoubtedly, however, the analogy has been an important one in the behaviourism or positivism which has sometimes appealed to social scientists. The razor can be seen in a characteristic destructive use in David Hume (1748, section XII, Part III). In a famous purple passage he invites us to ask of any volume: 'Does it contain any abstract reasoning concerning quantity or number? *No*. Does it contain any experimental reasoning concerning matter of fact and existence? *No. Commit it then to the flames: for it can contain nothing but sophistry and illusion.'*

A fourth and again a typically modern way of depicting the relationships between moral philosophy and the facts of the moral life can also be illustrated by David Hume, this time in his analogy with a *microscope* (1748, section VII, Part I). The philosopher takes the concepts used by the moral agent (or the social scientist) and looks at them through his philosophical microscope, thereby achieving a better understanding of their meaning. This is in some ways like the 'razor' approach because, although less destructive in intent, it often has had the same result of casting doubt on the validity of what cannot be empirically supported. Nevertheless, it is a view of the relationship between theory and practice which has been and still is philosophically influential.

A fifth analogy is that of *cartography*. The philosopher, it has been said, should be concerned with the logical mapping of concepts in a particular area, and so the moral philosopher will be concerned with the mapping of concepts in the moral and political life. This approach, as a supplement to Aristotle's, has a lot to be said for it. It suggests that the moral philosopher must himself learn from the moral agent or the practitioners of special subjects the procedures and types of

argument actually used, and it discourages philosophers from abstracting a concept from its context and examining it in terms of some artificially imported criterion of meaningfulness. Moral philosophers inspired by Wittgenstein were encouraged to see moral concepts as interdependent, like those of a game, and as part of a total way of life (Wittgenstein 1953).

It will be noted that of the five analogies the first two assign to moral philosophy some sort of normative function while the other three assign it an analytical function. In the period from 1945 to the present that distinction appeared in the literature as one between 'normative ethics' and 'meta-ethics'. Some philosophers even took the line that only the analytical approach was properly philosophy, and that moral philosophy should be entirely neutral on first-order moral questions. This doctrine went in conjunction with a second – that moral philosophy should be sharply distinguished from anything empirical – psychology, economics, sociology etc. These two doctrines – of the moral neutrality and non-empirical nature of moral philosophy – were at the root of a style of philosophy, known as linguistic analysis, which was the dominant one in Anglo-Saxon philosophy during the period 1945–65.

From the mid-1960s cracks began to appear in this solid front. Some philosophers questioned whether meta-ethics or analytical moral philosophy could be value neutral even if its practitioners tried. It was argued that apparently neutral, logical analysis of the language of morals presupposed a particular moral stance, liberalism, say. This debate was paralleled by the debate as to whether there could be a value-neutral social science. The conclusion that began to emerge – that meta-ethics cannot be morally neutral – had a liberating effect on practitioners of normative ethics, and the 1970s saw a resurgence of interest in normative questions. Moreover, the second pillar of the analytical approach weakened also, and there was a consequent development within professional philosophy of areas which had been out in the cold for several decades, such as political philosophy and the philosophy of education. It is true that some practitioners of philosophy of education (say) argued that they were just applying to other subject-matters the techniques of analytical philosophy, but the distinction between philosophy and empirical subject-matters was by now blurred, and new areas of interest to moral philosophers were opened up, such as social work ethics and medical ethics. At the time of writing, in the mid-1980s, the normative aspect to moral philosophy is if anything the dominant one, and philosophers produce books with titles like *What Sort of People Should There Be?* (Glover 1984), without any suggestion that they have strayed away from the central concerns of moral philosophy.

Let us now examine in more detail the typical problems and theories of 20th-century moral philosophy. It is convenient for expository purposes to accept the distinction between 'normative' and 'meta-ethical' theories. Normative theories are concerned with attempts to answer the questions, 'What makes right actions right?' or 'What makes actions duties?' The answers fall into two broad categories: the teleological, and the deontological. In other words, for some philosophers actions are right if they produce some sort of good state of affairs, while for others rightness is in some way intrinsic to the right action and independent of what the action might produce. In expounding teleological theories we must note a distinction, often overlooked, between consequentialist and non-consequentialist versions of teleology. A consequentialist sees the good to be brought about as being externally related to the right action; actions are right if they are instrumental in bringing about good states of affairs.

The most common and influential of these consequentialist theories is hedonistic utilitarianism, familiar in the slogan associated with J.S. Mill (1863): 'Seek the greatest happiness of the greatest number.' The theory has two components – a doctrine of right action (that actions are right if they produce the best possible consequences for the majority) and a doctrine of the good, or of the nature of these consequences (pleasure or happiness). The latter doctrine has been criticized on the grounds that not all pleasures are equally good, or that happiness is not the only thing good in itself. J.S. Mill was himself fully aware of the force of this sort of point and tried to defend the theory by introducing a

qualitative distinction between pleasures (Mill 1863). Whereas Bentham had maintained the consistent view that, if the amount of pleasure were the same, push-pin was as good as poetry, Mill tried to argue that some pleasures were *qualitatively* better than others. It is generally thought that Mill's argument is in itself circular, and also inconsistent with the rest of his theory. More recent hedonistic utilitarians have dropped the reference to pleasure or happiness and speak simply of people's 'interests' or 'preferences'. But it can still be argued that some things are better worth having than others regardless of whether or how much people might prefer them, or how much they might claim that these things were in their interests. This anti-hedonist argument, even if thought successful, need not lead to the abandonment of utilitarianism, but can (and historically did) lead to the development of non-hedonistic or 'ideal' forms of utilitarianism.

The utilitarian doctrine of right action, which is the crux of the theory, has been criticized on various grounds. First, it might be held to be radically unclear. The criterion refers to the best possible consequences 'for the majority'. But how is this to be interpreted? As referring to a small group (a family), or a community, or a race, or the whole of mankind? Does it include future persons? Does it include animals? The answers to these questions have important social and economic consequences, but there are no clear lines of guidance from the theory.

Second, supposing the theory could be stated so as to meet these criticisms of internal detail, we must still consider whether it can meet criticisms from the other main group of normative theories – the deontological ones. Theories of this type claim that certain types of action are just right or just obligatory, regardless of the consequences. Some deontological theories claim that no actions are right or are duties because of their consequences. In so far as Kant's moral philosophy is deontological he holds this position (Kant 1785). Other deontologists allow that some actions are right for utilitarian reasons and simply deny that all actions are right for utilitarian reasons (Ross 1939). But all deontologists would agree that duties of justice cannot be accomodated on any utilitarian scheme. For example, a deontologist would argue that the utilitarian idea of maximizing good seems to sanction the possibility of unequal distribution of good, and so could conflict with widespread and basic intuitions of equality or fairness, or *distributive justice*. Again, deontologists would argue that a utilitarian is committed to the punishment of the innocent if that would (as sometimes it might) maximize good. But this conflicts with our basic intuition of *retributive justice*. Yet again, the deontologist would stress that whereas the utilitarian is necessarily committed to the view that *duties of truth-telling* or *promise-keeping* are dependent on the maximization of good, the ordinary person's view is that such requirements seem to be moral duties regardless of the consequences.

Theories which stress the importance of *rights* can also be included under the umbrella of deontology. Indeed, theories stressing rights and deontological theories stressing duties both arose historically from doctrines of natural law going back to the Greeks. Theories of rights became popular in the 17th and 18th centuries as doctrines of 'natural rights' or the 'rights of man', and have again become popular in this half of the 20th century in the vocabulary of human rights. Many moral causes are supported by invoking human rights. Theories of rights have this in common with all deontological theories that they insist that rights can be held by individuals regardless of the interests of the majority.

One of the interests in normative moral philosophy during the period 1960–75 was the rise and fall of utilitarian attempts to come to terms with deontological criticisms of their theory. Those attempts were expressed in a theory known as 'rule-utilitarianism' (as distinct from 'act-utilitarianism'). Rule-utilitarians distinguished rules of two sorts. The first were sometimes called 'regulative', such as 'People ought not to lose their tempers'. These were said to be rough guides, rules of thumb, to the likely best consequences of individual actions, and had no other force. The second were sometimes called 'constitutive' or rules which define a 'practice', such as 'One ought to keep one's promise', or 'One ought to tell the truth'. Property rules and rules of justice

would also come into this category. The point to notice about the second category, according to the rule-utilitarian, is that we must distinguish two questions which were not distinguished by either the act-utilitarian or the deontologist: why ought we to perform certain individual actions? and, why ought we to agree to having certain sorts of rule? The deontologist correctly answers the first question by saying that we ought to keep our promise (say) simply because a promise is a promise, but he cannot on his theory explain how we can justify accepting the whole practices of promising, or of owning property etc. The act-utilitarian, on the other hand, who does see the importance of good consequences for mortality, does not see that a 'practice' will be undermined and rendered useless if people decide for themselves on each occasion whether or not the best consequences will come about by their actions. The solution, according to the rule-utilitarian, is to insist with the deontologist that we keep to the rule because it is a rule, but then to go on to say that the existence of the rule can be justified, if it can, to the extent that its operation as a whole produces the best possible consequences for the majority. This theory is associated with the name of John Rawls (Rawls 1955), but it can be found as powerfully expressed in David Hume (Hume 1739).

Rule-utilitarianism was much criticized during the 1970s. The usual line was to note that it tends to inconsistency. The rule-utilitarian tells us that we should keep to a rule, say that we should keep a promise or observe a right, even when we know for sure that doing so will not maximize the best consequences in the given case, for otherwise we shall undermine the rule or the 'practice'. But if the promise has been made in secret, as promises often are, then it is not clear how we would be undermining the practice if we broke the promise. Many other examples can be used to make the point that the attempt to combine the views of the act-utilitarian with those of the deontologist results in inconsistency.

There is a third sort of criticism which can be made of any kind of utilitarianism. We can put it in terms of rights, although it can be put in other ways and does not depend on accepting deontology. If A injuries B in some way – assaults him or slanders him – then this may have bad consequences for the majority. For example, it might encourage similar sorts of bad behaviour. But that is not what is mainly wrong with A's action. What A has done is wrong, if it is wrong, not because of the effect of his action on the *majority* but simply because he has injured B. If we put this in terms of rights we can say that B's rights have been violated.

There is a fourth sort of criticism of utilitarianism, which refers to the moral position of the agent of the action, as distinct from the recipient. The concept often invoked is 'integrity'. For example, let us suppose that A is on a bus party and is captured by bandits. The bandit chief (a man of honour) promises to release A plus all the other members of the bus party provided A agrees to shoot any one member of it. If A does not agree then the bandit chief will himself shoot everybody. Now on utilitarian terms there is no doubt that A should agree to shoot one person if the others can go free; but it is at the very least not morally unintelligible if A should think he ought to refuse, on the grounds that *he* ought not to kill someone himself even at the cost of someone else's killing far more people (Smart and Williams 1973).

In general terms, what is lacking in utilitarianism is some sort of appreciation of the essential connection between morality and the nature of a person. The theory is at its most plausible if it is considered as an administrator's or legislator's theory, but it lacks any grasp of the inward, or personal, or face-to-face, aspects or morality.

A possible way forward is to return to the distinction between consequentialist and non-consequentialist teleology and to consider the latter. The Greeks provided one sort of non-consequentialist teleology, and the Idealists of the 19th century provided a slightly different sort. Common to both was the idea of morality as being an expression of essential aspects of human nature and social life. For the utilitarians, morality was simply a device, instrumental in producing a harmonious society; and human beings were conceived as being simply consumers of happiness. To put it in another way, the

consequentialist teleology of utilitarianism is external to the self. But it is possible to see morality in terms of an internal teleology. An analogy might help here. When a house is built the lorries which bring the materials, the cement-mixers, the scaffolding, are all necessary means to the final product, which is the house. They have a value in their use, but are externally related to the product and have no part in it when it is finished. By contrast, the bricks and cement and tiles, while they too are necessary means for the creation of the final product, are also internally related to it; indeed, they are part of it, and it displays its character through them. In an analogous way, I suggest, morality should be seen not just as a means necessary for the good life for man, but as itself an expression of it. Human beings have a moral nature as well as a capacity for happiness, and it is this that the utilitarians ignored. The deontologists were correct in seeing that morality cannot be justified simply in terms of the good things it might produce, but they too (although we must except Kant) lacked a grasp of the connection between morality and the self. Insights into this connection are, curiously, shown by J.S. Mill. He stresses the importance of the 'flourishing' of the self, a Greek idea which he sees in non-hedonistic terms. As he puts it, 'It really is of importance not only what men do, but also what manner of men they are that do it' (Mill 1859, chapter 3). In other words, morality must be seen not simply as a technical device, but as a humane practice.

Let us move from normative questions to meta-ethical ones, bearing in mind that they too presuppose views about the nature of man and are far from being morally neutral. Within the broad area of meta-ethics three types of question are often discussed and not always distinguished. First, there are questions of what is sometimes called the 'logic of moral language'. These are often thought to be questions of the meanings of moral words. Secondly, there are questions of moral epistemology, of how we can be said to know the difference between right and wrong, or whether the vocabulary of knowing is appropriate. Thirdly, there are questions of the metaphysics of morality. Is it just an expression of human emotion or is it somehow part of the fabric of things? The questions are all interconnected, in that answers to one of the sets of questions have implications for the others. Let us begin with the logical questions, but note that they will slide into the others.

Three possibilities have been discussed this century: that moral judgements are statements of fact, that they are expressions of emotion, and that they are commands or prescriptions.

The view that they are statements of natural fact has been called 'naturalism', for it is along the lines that moral judgements are in some way 'reducible to' or 'analysable in terms of' some sort of fact, such as 'what gives pleasure'. G.E. Moore (1903) went as far as to say that philosophers who held this view had committed a logical fallacy, which he named 'the naturalistic fallacy'.

This was the alleged fallacy of defining 'good' in terms of something other than itself, and Moore thought that the fallacious nature of this definition could be brought out by what came to be known as the 'open question' test. For any attempted definition of 'good' in terms of properties x, y, z, it is possible to say, 'This is x, y, z, but is it good?', and the fact that this question remained open (or at least meaningful) was thought to indicate that the definition could not be correct. Moore concluded that 'good' did not name a natural property, but was indefinable and irreducible (although other moral words like 'right' could be defined as 'what realizes good'). He inferred from his conclusion, that 'good' does not name any natural property, that it must name a *non-natural* property. This was said to be in some ways like a natural property such as 'yellow', but in other ways quite different. In Moore, then, we find a logical view – that 'good' is the name of an indefinable property – which commits us to a metaphysical view – for since the property is non-natural it cannot be understood by any of the sciences, but it is still part of the fabric of things. This metaphysical view in turn commits us to an epistemological view, for since the property is non-natural it cannot be known by any of the senses, so must be known by an 'intuition'.

All three of these aspects of Moore's position were heavily criticized. For example, some

philosophers argued that Moore had an inadequate view of language, since he thought that adjectives must always name properties. It was argued that 'good' does not name any sort of property, but rather commends, or expresses favourable emotion, or a pro-attitude, to something. This argument was based both on a philosophy of language, and also on a revived awareness of the action-guiding nature of moral judgements. The action-guiding nature of moral judgements is not explained if they are taken to state facts or name properties, but seems to be better explained if they are taken to express attitudes.

Out of this came the theory known as 'emotivism', which was immensely influential in the 1940s and 1950s. For the emotivist, moral sentences do not primarily state facts; they express attitudes. The expression of attitudes in moral utterances has a 'magnetic' effect – the hearer is moved to act by them. To support a moral utterance with a reason is to mention a (natural) fact which will causally influence an attitude (Stevenson 1937). This view also was criticized in many ways. For instance, we often make moral judgements with no intention of influencing other people's attitudes, as when someone says to a vegetarian that eating meat is wrong. Again, it is not clear how emotivism can cope with moral doubt, as when I *wonder* whether I ought to do X or not. But, above all, emotivism does not seem to do justice to the apparent rationality of morality. We argue with people and give reasons for our positions, and we think other people are mistaken. But how can a person be mistaken if his moral view is just a matter of causally induced emotion?

In an effort to regain the rationality of morality, while preserving the practicality stressed by emotivism, R.M. Hare developed a theory which became known as 'prescriptivism' (Hare 1952, 1963). The prescriptivist line on practicality was that the moral agent chooses his ultimate moral principles, and to choose a principle is to commit oneself to what the principle enjoins. 'I ought to do x' is seen as being like a command expressed to oneself, or like a firm statement of intention. The *practical force* of moral judgements then is expressed *via* the logical thesis that to assent to a moral judgement is to be committed to acting in terms of it; if one does not so act then one does not hold the principle. Such a position gave rise to many problems over weakness of will. Can I not still be said to hold a principle even if sometimes I weakly and blindly, or even perversely and deliberately, act against it?

Hare was emphatic that the rationality of moral judgements could not be explained by showing them as deducible from any set of factual premises – a moral 'ought' logically cannot be deduced from any set of 'is' propositions. This argument which originated in Hume, was used as an apparent trump card against naturalists. It was Hare's version of Moore's 'naturalistic fallacy'. The rationality of moral judgements was said by Hare rather to be expressed *via* the idea of the 'universalizability' of moral judgements. If I say that I ought or ought not to do X then I am logically committed to saying that anyone in a similar position ought or ought not to do likewise. But this provides only a thin account of rationality. The account is correct in insisting that singular 'oughts' must be capable of being made universal by being connected with general rules, but those who see morality as rational are claiming to see more than consistency in it. For example, if I say that I ought not to walk on the cracks of the pavement then I am certainly committed to the rule that no one ought to, but the rule itself does not sound justifiable. Hare thinks that in practice people will not prescribe rules which are not in their self-interest, and dubs those who do 'fanatics'. But he cannot show that the fanatic is mistaken; all he can do is to weaken the appeal of fanaticism by invoking the utilitarian's generalized self-interest.

Dissatisfied with prescriptivism and its many difficulties, and with the aim of restoring rationality to morality in a sense stronger than consistency, some philosophers such as Philippa Foot (1958) or G.J. Warnock (1971) revived naturalism. What these forms of naturalism had in common was the attempt to show how morality was logically connected with concepts of harm and benefit. This certainly brought back objectivity and rationality to morality. To use Warnock's example, to triumph over one's enemies may be a splendid thing, but it must, objectively, be

morally wrong, granted the premise that morality is logically connected with the concepts of harm and benefit plus some incontestable empirical premises about human reactions.

But neo-naturalism still has difficulties with the practical force of moral judgements. How do natural facts provide reasons for action unless we assume dubious premises such as that we all desire each other's benefit, or that my own benefit is furthered by that of others?. The main problem with this neo-naturalism, however, is that either it makes morality too narrow in scope – confining it to matters of human harm and benefit – or it makes the conceptions of harm and benefit so wide that they are emptied of all meaning. For example, two consenting single adults who have a sexual encounter and who take contraceptive precautions cannot be said to be harming anyone (if 'harm' is to mean anything). Yet some people might condemn what they do on moral grounds. Now, whether or not we agree with this condemnation, we cannot rule it out on logical grounds as not being a moral condemnation. But this is precisely what we are committed to if morality is narrowly defined in terms of human harm and benefit. And what about harm and benefit to animals?

It should be noted that the moral philosophers after Moore – the emotivists, the prescriptivists and the neo-naturalists – all have in common that they assume a *metaphysical* naturalism. In other words, they all assume that the whole phenomenon of morality can be explained in terms of some combination of psychology, economics and sociology (plus, nowadays, socio-biology). At the moment, however, there are faint signs of a revival of interest in non-naturalistic views of morality. This has been brought about partly by the rediscovery of Hegel, and partly by a revival of Natural Law studies. At any rate, it provides a welcome shake to the kaleidoscope of moral philosophy.

No account of the subject would be complete without some reference to Existentialism. Continental philosophy has never been part of the main stream of Anglo-Saxon philosophy, but many of the concepts of Existentialism have been accepted by moral philosophers. For example, the idea of 'bad faith', of pretending to oneself that one is determined, that one has no choices so must just accept a way of life, was one discussed by J.-P. Sartre (1943) in a series of illuminating examples. Indeed, one beneficial influence which Existentialism exerted on Anglo-Saxon philosophy was *via* its more dramatic examples. This encouraged moral philosophers to abandon the thin examples offered by duties of returning library books or the rules of cricket in favour of more extended and full-blooded ones.

This use of more realistic examples, plus the desire to be applied to other subject-matters (such as medicine), plus the availability of more sophisticated logical techniques, all make moral philosophy a more worthwhile subject than it was 40 years ago.

See Also

- ▶ Enlightenment, Scottish
- ▶ Entitlements
- ▶ Hedonism
- ▶ Philosophy and Economics
- ▶ Self-interest
- ▶ Solidarity
- ▶ Utilitarianism
- ▶ Utopias

Bibliography

Aristotle. 1954. *Nicomachean ethics*. Trans. Sir David Ross. London: Oxford University Press.

Foot, P. 1958. Moral beliefs. In *Aristotelian society proceedings*. Collected in P. Foot, *Virtues and vices*. Oxford: Blackwell, 1978.

Glover, J. 1984. *What sort of people should there be?* Harmondsworth: Pelican Books.

Hare, R.M. 1952. *The language of morals*. Oxford: Clarendon Press.

Hare, R.M. 1963. *Freedom and reason*. Oxford: Clarendon Press.

Hume, D. 1739. In *A treatise of human nature*, ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1896, 2nd ed, 1978.

Hume, D. 1748. In *An enquiry concerning human understanding*, 2nd ed, ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1902. Revised by P.H. Nidditch, 1975.

Kant, I. 1785. *Groundwork of the metaphysic of morals*. Trans. H.J. Paton as *The moral law*. London: Hutchinson's University Library, 1948.

Mill, J.S. 1859a. In *Essay on liberty*, ed. M. Warnock. London: Collins, 1962.
Mill, J.S. 1859b. In *Utilitarianism*, ed. M. Warnock. London: Collins, 1962.
Moore, G.E. 1903. *Principia ethica*. Cambridge: Cambridge University Press.
Plato 1888. *Republic*. Trans. Benjamin Jowett, 3rd ed, revised. Oxford: Clarendon Press. Reprinted. New York: Sphere Books, 1970.
Rawls, J. 1955. Two concepts of rules. *Philosophical Review* 64: 3–32.
Ross, W.D. 1939. *Foundations of ethics*. Oxford: Clarendon Press.
Sartre, J.-P. 1943. *Being and nothingness*. Trans. Hazel Barnes. London: Methuen, 1957.
Smart, J.J.C., and B. Williams. 1973. *Utilitarianism: For and against*. Cambridge: Cambridge University Press.
Stevenson, C.L. 1937. The emotive meaning of ethical terms. *Mind* 46(181): 14–31.
Warnock, G.J. 1971. *The object of morality.* London: Methuen.
Wittgenstein, L. 1953. *Philosophical investigations*. Trans. G.E.M. Anscombe. Oxford: Blackwell.

Morgenstern, Oskar (1902–1977)

Martin Shubik

Keywords

Game theory; Morgenstern, O.; Perfect foresight; Prediction; Rational behaviour; von Neumann, J.

JEL Classifications

B31

Morgenstern was born in Goerlitz, Silesia, on 24 January 1902. He died on 26 July 1977 at his home in Princeton, New Jersey. The two main intellectual centres of his life were Vienna and Princeton. In each case the source of his intellectual stimulation was not primarily the university but institutions such as the Wienerkreis of Moritz Schlick in Vienna, where he counted among his friends Karl Popper, Kurt Gödel and Karl Schlesinger, and the Institute for Advanced Study at Princeton. He obtained his doctorate in 1925 from the University of Vienna, where he was greatly influenced by Karl Menger and the writings of Eugen Böhm-Bawerk.

Morgenstern's first major work, *Wirtschaftsprognose* (1928), which was published in Vienna, served as his Habilitation thesis leading to his appointment as a *privatdozent* at the University of Vienna in 1929. In this book he began to consider the difficulties and paradoxes inherent in economic prediction, being particularly concerned with prediction where the action of a few powerful individuals could influence the outcome. He illustrated some of these difficulties with the example of Sherlock Holmes's pursuit of Professor Moriarty (an example repeated in the *Theory of Games,* 1944).

He became a professor at the University of Vienna in 1935, and in the same year published in the *Zeitschrift für* Nationalökonomie(of which he was managing editor) an article on fundamental difficulties with the assumption of perfect foresight in the study of economic equilibrium. It was then that the mathematician Edward Čech noted that the problems raised by Morgenstern were related to those treated by von Neumann in his article 'Zur Theorie der Gesellschaftspiele', published in 1928.

Morgenstern did not have the opportunity to meet von Neumann until somewhat later. They both recalled meeting at the Nassau Inn in Princeton on 1 February 1939, although each believed that they had met once before. They became close friends and remained so until von Neumann's death on 8 February 1957.

In Vienna, Morgenstern was also director of the Austrian Institute for Business Cycle Research (1931–8) where he employed Abraham Wald, whom he later helped go to the United States. In 1938, due to his opposition to the Nazis, Morgenstern was dismissed from the University of Vienna as 'politically unbearable' and he accepted an offer from Princeton, to some extent because of the presence of von Neumann at the Institute for Advanced Study. Their close collaboration resulted in the publication in 1944 of their book, *The Theory of Games and Economic Behavior.* This major work contained a radical reconceptualization of the basic problems of

competition and collaboration as a game of strategy among several agents, as well as an important novel approach to utility theory (presented in detail in the second edition, 1947).

Both Morgenstern and von Neumann were well aware of the limitations of their great work. They stressed that they were beginning by offering a sound basis for a static theory of conscious individually rational economic behaviour and that the history of science indicated that a dynamic theory might be considerably different. They warned against premature generalization.

In his years at Princeton from 1938 until his retirement in 1970, Morgenstern encouraged the work of a distinguished roster of younger scholars in game theory and combinatoric methods. This was feasible primarily through the strength of the Mathematics Department and its connections with the Institute. There was little interest in the subject in the Department of Economics at the time. The ideas of the *Theory of Games* were so radical that they have taken many years to permeate the social sciences. Even at the time of his death many in the economics profession were sceptical of or indifferent to its contributions.

Although his work on the theory of games was undoubtedly Morgenstern's greatest contribution and collaboration, his interests were wide-ranging. His two books, *On the Accuracy of Economic Observations* (1950), and *Predictability of Stock Market Prices* (1970), written jointly with Clive W. Granger, indicate these interests. He was also concerned with matters of national defence and in 1959 published *The Question of National Defense*.

In 1959 he was one of the founders of Mathematica, a highly successful and sophisticated consulting firm, and served as Chairman of the Board. After retiring from Princeton he was Distinguished Professor at New York University until his death.

Selected Works

1928. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten. Vienna: Julius Springer.*

1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* 6(3): 337–357.

1944. (With von Neumann J.). *Theory of games and economic behavior*, 2nd edn. Princeton: Princeton University Press, 1947.

1950. *On the accuracy of economic observations.* Princeton: Princeton University Press.

1959. *The question of national defense. New York: Random House.*

1970. (With Granger C.W.J). *Predictability of stock market prices.* Lexington: Heath Lexington Books.

Morishima, Michio (1923–2004)

Meghnad Desai

Abstract

Morishima's contribution to economic theory was in tackling questions of equilibrium and dynamics with and without money, with heterogenous capital and in a multisectoral framework. He tried to synthesize and answer questions raised by Ricardo, Marx, Walras, Wicksell, Keynes and Schumpeter. His work was influenced by von Neumann's model and Hicks's style of theorizing.

Keywords

Capital accumulation; Capital controversy; Credit creation; Deflation; Econometric Society; Effective demand; Excess demand and supply; Exploitation; False trading; Growth paths; Heterogeneous capital; Hicks, J. R; Homogeneity postulate; Inflation; Innovations; Investment functions; Involuntary unemployment; IS–LM models; Keynes, J. M; Marx, K. H; Marx–von Neumann model; Microfoundations; Morishima, M; Natural rate of interest; Nominal demand; Nominal rate of interest; Quantity theory of money; Ricardo, D; Saving and investment; Say's Law;

Schumpeter, J. A; Stability of equilibrium; Tâtonnement; Turnpikes; Von Neumann, J; Walras, L; Walras–Leontieff model; Wicksell, J. G. K

JEL Classification
B31

Michio Morishima was one of the most distinguished economic theorists of his generation. He taught in Japan at Kyoto and Osaka Universities, and in the UK he was the Keynes Visiting Professor at the University of Essex 1969–70 and Professor of Economics, later the John Hicks Professor of Economics, at the London School of Economics 1970–84 and Emeritus Professor for the rest of his life. He was awarded the Order of Culture [Bunka Kunsho] of Japan by the Emperor in 1976, a Fellowship of the British Academy in 1981 and an Honorary Fellowship of the LSE upon his retirement. Morishima became the first Japanese to be the President of the Econometric Society in 1965. He died aged 80 on 13 July 2004, leaving behind his wife Yoko and two sons and a daughter.

Morishima's work encompasses general equilibrium theory with heterogeneous capital, growth and money, as part of a coherent attempt to tackle one of the most intractable problems in economic theory, namely, the construction of an adequate theory of a dynamic growing economy with heterogeneous capital and money as well as credit or, to put it another way, a theory of how the capitalist system works.

Morishima's Ph. D. thesis at Kyoto University (published in Japanese in 1950 and in English in 1996 under the title *Dynamic Economic Theory*) dealt with stability of equilibrium. The standard (Hicksian) theory says that if the market starts out at a price away from the equilibrium given by the intersection of the demand and supply curves, then the price must change until the equilibrium point is reached. But how? Walrasians posit an auctioneer who would call out prices and register demands and supplies at each price. No trades are made until the auctioneer is satisfied that demands and supplies balance, that is, no false trading. The corollary of a no false trading equilibrium is that there can be never be involuntary unemployment, raising the issue of the consistency of micro and macro theories with each other.

Morishima prefers the case in which trading takes place at each price, but the price changes if, at that price, after transactions are closed, there is excess supply or demand. This would be a non-tâtonnement process, where some traders may buy (sell) at a price higher (lower) than the equilibrium price. He does not, however, develop this any further in the thesis but asks: are we exploring the path of convergence of the 'groping' prices, that is, virtual prices at which no trades are carried out and hence within 'the market day', or are we talking of the path of equilibrium prices arrived at, at the end of the tâtonnement in each market day from one day to the next?

Within the Hicksian week, the groping process traces out a path of virtual prices which converge to equilibrium under certain well-known conditions. But what of the sequence over several weeks of the equilibrium price? What are the dynamics of the path itself? It is this question that Morishima poses in *Dynamic Economic Theory* and pursues over his entire career. It is obviously connected to the stability of a growth path, since the path of income is analogous to the path of equilibrium prices. Morishima's discussion of growth paths was therefore always concerned not only with the quantity variables such as income and the stock of capital but also prices and interest rates.

Morishima's first book in English, *Equilibrium, Stability and Growth* (1964), tried to integrate Walras into the growth story, which had not hitherto been attempted, and also gave prominence to Marx's work on accumulation at the same time. Morishima constructed Walras–Leontieff and Marx–von Neumann models, which are pioneer efforts. *Equilibrium, Stability and Growth* is growth-oriented with an emphasis on linear technology and balanced maximal growth paths with fixed coefficients. But there is also a chapter on a spectrum of techniques. This is Morishima's response to the then ongoing capital controversy between Cambridge England and Cambridge Massachusetts.

Very soon after *Equilibrium, Stability and Growth* was published, Morishima came out

with his most ambitious work to date, Theory of Economic Growth (1969). Here Morishima sets out a rigorous multisectoral framework – the von Neumann model – and integrates Walras as well as Hicks into this framework. Prices are solved out along with quantities throughout. Turnpikes are discussed under various assumptions. But Morishima also deals with the issue of the optimality of the maximal growth paths.

Morishima was not happy with *Theory of Economic Growth*. Thus started his long detour via Marx, Walras and Ricardo, until he could come back to his major concern. Morishima's book *Marx's Economics* (1973) deals with the statics and dynamics of Marx's growth and exploitation theory and tackled joint production with innovative insights. It shows that labour values can be used to tackle the aggregation problem for heterogeneous capital.

The crucial next step is provided by Walras. Most economists think that Walras provided consistent microfoundations for a full employment–all markets clearing theory of the macroeconomy. Morishima had a different Walras in his 1977 book with the intriguing title *Walras' Economics: A Pure Theory of Capital and Money.* Morishima's purpose in the book is to see whether he can exploit Walras's work to provide the microfoundations of Keynesian macroeconomics. He focuses on the contrast between nominal demands (neoclassical) and effective demands (Keynesian) as well as the alternative hypotheses that investments adjust to savings (neoclassical) and that investments are prior and saving adjust (Keynesian). Walras's entrepreneurs have no income; they work on altruistic principles. Morishima adjusts Walras's investment function as well as giving entrepreneurs an income (profits) which makes the model closer to real capitalism. But he also shows why one needs a theory of accumulation and growth, that is, a story with time and future in it, in order to have a rationale for holding money in a Walrasian world. In a static general equilibrium, money can, and does, play no role.

The heart of Morishima's book *Ricardo's Economics* (1989) is in the final section entitled 'Three Paradigms Compared'. Say's Law is at issue. Ricardo established Say's Law as a dominant mode of theorizing. Usual departures from Say's Law involve a non-trivial role for money and/or a growth process via an active investment function. Ricardo had neither and so could subscribe to Say's Law. Marx had both but his investment function was very restrictive and made no use of money or credit. Walras had money towards the end of *Elements* but his growth theory lacked an investment function which led the way for savings to adjust to investment. Keynes of course had money and investment functions, but he did not spell out the microfoundations. Growth is not sufficient to justify a violation of Say's Laws; money or an investment function which has a role for entrepreneurs to respond to uncertainty is required.

In *Ricardo's Economics* a model is set up in which excess demand and supply for labour and capital are modelled in a simple diagram (1989, fig. 6, p. 218). Here, around an equilibrium point, zones of excess supply and demand for the two factors are mapped out. Morishima's axes are the real wage and the output capital ratio. Within the same general model all the three paradigms are embedded. Again, the investment function turns out to be the crucial relationship for the Anti-Say's Law result that Keynes established.

Capital and Credit: A New Formulation of General Equilibrium Theory (1992) brings together all the major themes of money, heterogeneous capital, underemployment equilibria and growth. The major innovation in *Capital and Credit* is that banks play a crucial role in financing production. This is Schumpeter rather than Keynes. While in Keynes's scheme entrepreneurs may underinvest because of expectations or a low marginal efficiency of capital relative to the rate of interest, Schumpeter allows for overshooting of credit creation by bankers. Thus, inflation as well as underemployment is possible.

Capital and Credit is therefore concerned with innovations and their financing and monetary disequilibrium. The economy is split into Say's Law and Anti-Say's Law activities. There is a scope for Anti-Say's Law if production is financed by credit, and this of course requires that it is not instantaneous but has an input–output lag.

With instantaneous production and investment adjusting to savings, Say's Law is confirmed. But in any realistic capitalist economy it breaks down due to the presence of credit. The amount of credit determines activity in the Anti-Say's Law sector (manufacturing industry, in other words), and this, via the multiplier, determines the overall levels of activity and employment. This need not be full employment. The separation of the economy as between relative prices determined by demand/supply and absolute prices as determined by money –'the classical dichotomy' is no longer valid. It is only by omitting banks and the financial requirements for production that the dichotomy is sustained.

In the last chapter, on 'Monetary Disequilibrium', Wicksell's cumulative process is examined from the point of view of von Neumann. The real system establishes the rate of profits (= rate of growth), but it leaves the price level indeterminate. Credit creation by bankers determines the nominal level of interest with the natural rate given by the real system. Then the monetary side determines the price level by the intersection of the money demand function and the real growth rate. But it is not a stable equilibrium. It is a kind of IS–LM model, but with its axes as interest rate and price level rather than income.

So we now enter a new development in monetary and growth theory. If the economy is growing and/or if the natural rate is a variable, then we need to extend Wicksell's analysis, which assumed a constant natural interest rate. But the natural rate may be above or below the von Neumann rate, and if the natural rate is also variable then the gap between the natural and the money rate is variable over the cycle. Thus, if the natural rate is above the money rate and the von Neumann rate, then inflation follows, but that may reduce the natural rate. If it then crosses over to being below the money rate, deflation follows and the natural rate may approach the von Neumann rate from above. Prices keep falling, and the economy may converge to the von Neumann rate.

In the converse case, the economy starts off with the natural rate below the money rate and below the von Neumann rate, and then deflation comes first as the natural rate approaches the von Neumann rate from below. Once it crosses over the constant money rate, then inflation follows and the economy approaches the von Neumann rate in an explosive inflationary situation.

This is the most sophisticated discussion of money and growth in the classical Wicksell framework. A variable natural rate is seldom modelled, and the deflation–inflation cycles enrich the Wicksell model greatly. But we are still in the world of Say's Law. What happens if we break away from it? The shortage of credit will restrict the economy below full employment, as Keynes envisaged, and abundance of credit will start off an inflationary growth process, as Schumpeter said. This then is the climax of the entire edifice of Morishima's work. He can now combine Anti-Say's Law with credit and disequilibrium. Credit creation determines the natural rate via the Anti-Say's Law sector, which is often the most innovative and dynamic. Morishima can then tackle the classical dichotomy.

This is the homogeneity postulate whereby nominal variables cannot have real effects and so money must be a veil. But the homogeneity postulate requires that a monetary shock be evenly spread across all agents. It also requires that the elasticity of demand with respect to money balances be identical across all agents. Morishima shows in the final pages of *Capital and Credit* that neither of these assumptions is likely to be fulfilled in a monetary economy. Agents including households and firms and the Anti-Say's Law firms are much more credit-sensitive than other firms, for one thing. And if the homogeneity postulate falls, so does the quantity theory. The challenge of integrating money and growth with general equilibrium but without Say's Law has been accomplished. There is much more to be gained from a careful study of these writings and one can only hope that future scholars will mine the rich source of theoretical insights in the decades to come.

See Also

► Capitalism
► Dynamic Models with Non-clearing Markets

- ▶ Growth Models, Multisector
- ▶ Say's Law

Selected Works

1950. *Degaakuteki Keizai Riron*. Tokyo: Kobundo.
1964. *Equilibrium, stability and growth*. Oxford: Clarendon Press.
1969. *Theory of economic growth*. Oxford: Clarendon Press.
1973. *Marx's economics: A dual theory of value and growth*. Cambridge: Cambridge University Press.
1973. (ed.) *Theory of demand: Real and monetary*. New York: McGraw Hill.
1976. *Economic theory of modern society*. Cambridge: Cambridge University Press.
1976. *Walras' economics: A pure theory of capital and money*. Cambridge: Cambridge University Press.
1978. (With G. Catephores.) *Value, exploitation and growth*. London: McGraw Hill.
1986. *The economics of industrial society*. Cambridge: Cambridge University Press.
1989. *Ricardo's economics: A general equilibrium theory of distribution and growth*. Cambridge: Cambridge University Press.
1992. *Capital and credit: A new formulation of general equilibrium theory*. Cambridge: Cambridge University Press.
1996. *Dynamic economic theory [an English translation of Degaakuteki Keizai Riron*, 1950, with additional articles]. Cambridge: Cambridge University Press.

Mortality

James W. Vaupel, Kristín G. von Kistowski and Roland Rau

Abstract

Mortality is a demographic component that contributes to shaping the size, structure, and dynamics of populations. Life expectancy has been rising remarkably in the more developed countries since the 19th century and the process of rising life expectancy also has begun in most of the less developed countries. Increases in adult life expectancy and declines in birth rates result in aging societies. Survival is increasing as a result of progress in economic development, social improvements, and advances in medicine. However, death rates vary significantly in different parts of the world and are particularly high in sub-Saharan Africa.

Keywords

Biodemography; Central death rate; Demography; Fertility; Gompertz law of mortality; Health care; Innovations; Life expectancy; Marriage and divorce; Migration; Mortality; Pensions; Population aging; Standard of living; Time-series analysis

JEL Classifications

J10

Mortality is one of the three demographic components that shape the size, structure, and dynamics of populations; the other two are fertility and migration. Death rates have declined remarkably in modern times. The populations of the more developed countries have been aging for more than 100 years and the process of rising life expectancy also has begun in most of the less developed countries. Survival is increasing as a result of progress in economic development, social improvements, and advances in medicine. Mortality has been falling steadily especially in wealthier, economically advanced countries and has continued to do so during the second half of the 20th century and after, particularly at higher ages. We are getting older and the number of the elderly is increasing in most countries.

While the reduction in human mortality can be considered one of the greatest achievements of modern civilization, rising longevity and the increasing number of elderly will pose major challenges to health care and social security systems.

Declines in birth rates and increases in adult life-expectancy result in aging societies. These demographic changes will impact the life-course decisions of individuals, social interaction, economic development, and policy reforms in the countries involved.

Rising life expectancy has globally been a widespread phenomenon, but mortality differentials remain. Death rates vary significantly in different parts of the world and are particularly high in sub-Saharan Africa by global standards. Mortality conditions have changed throughout history and vary among and within populations. Death rates differ according to the country of origin, place of residence, sex, socioeconomic status, level of education and marital status.

Mortality and Life Expectancy

Various indicators exist to measure mortality. Two of the indicators most often used and cited are the central death rate and life expectancy. The former, which is age- specific and time-specific, is defined as the number of deaths occurring at a given age during a given year, divided by the mean population of that age and year.

Life expectancy is an estimate of average age at death under current death rates. It is calculated by imposing the age-specific death rates of the respective year on a hypothetical cohort of newborns. In 2004, Japan reached the highest female life expectancy (85.59 years) ever obtained by a country. Lowest life expectancy is generally recorded in sub-Saharan Africa. An example is Zimbabwe, a country that in 2004 suffered the world's lowest life expectancy, 34 years for men and 37 years for women, according to WHO (2006). The United Nations estimated worldwide life expectancy for 2000–5 at 67.7 and 63.2 for women and men, respectively (United Nations 2005).

Remaining life expectancy at age x is usually denoted as e_x and ${}^{e^0}{}_x$. A value of $x = 0$ leads to the most often published indicator, 'life expectancy at birth'. Note that 'life expectancy' for a given year is based on a hypothetical cohort. Only if death rates are not changing can the average newborn be expected to live the number of years indicated by life expectancy. If age-specific mortality continues to decrease – as was the case in many developed countries during recent decades – then the actual average age at death of a birth cohort would be higher than the one estimated for the hypothetical cohort.

Age Trajectories of Human Mortality and the Gompertz Law of Mortality

As individuals age, they tend to suffer an increasing loss of physical function and greater susceptibility to disease and injury. Benjamin Gompertz, a British actuary, described in 1825 the gradual increase in mortality rates with age, using an exponential curve, today known as the 'Gompertz law of mortality'. The model implies that there is a constant rate of increase in the age-specific mortality of adult populations; for many populations this rate of increase is about ten per cent per year. The Gompertz model fits human mortality rates well for adults aged 30–85 in most modern populations with high life expectancies.

The overall age trajectory of human mortality is roughly U-shaped. Mortality is high immediately after birth. During infancy it decreases rapidly with age to reach a minimum between the ages of 10 and 15. Thereafter, the risk of dying rises more or less exponentially according to the Gompertz law of mortality, with some excess mortality among young adults. A rise in mortality during early adulthood is often referred to as 'accident hump', as it is mainly caused by accidents in many modern populations (Heligman and Pollard 1980). Especially in industrialized countries, this hump is more pronounced for men than for women. The hazards associated with being a woman of childbearing age have been greatly reduced in developed countries, but those connected with the transition to manhood are still substantial. Maternal mortality, in contrast, is confined almost exclusively to developing countries. Among women worldwide, those in sub-Saharan Africa are at highest risk of dying during pregnancy and at childbirth: The lifetime risk was estimated by WHO at 1 in 16 in 2002;

this compares to a risk of 1 in 2,500 in the United States (WHO 2004).

Most deaths in developed countries today are concentrated at older ages. Death rates at older ages have, however, declined markedly during the second half of the 20th century. Furthermore, after age 80 death rates rise more slowly than predicted by the Gompertz exponential formula, and may roughly level off around age 110, albeit at the high level of about 50 per cent mortality per year (Thatcher et al. 1998; Robine and Vaupel 2002).

Rising Life Expectancy in Industrialized Countries

The rise in life expectancy is one of the great achievements of modern times. In the countries with the highest levels, female life expectancy has been rising for 160 years at a steady pace of almost three months per year (Oeppen and Vaupel 2002). The four-decade increase in best-practice life expectancy is so extraordinarily linear that it may be the most remarkable regularity of mass endeavour observed. On average, women live longer than men, but record life expectancy has also risen linearly for men since 1840, albeit a little more slowly than for women. The improvements in survival leading to the linear climb in record life expectancy result from the intricate interplay of advances in income, salubrity, nutrition, education, sanitation and, in recent decades, medicine (Riley 2001).

When we look at individual countries, gains in life expectancy have not progressed as linearly. The gap between the record level and the national level can be regarded as a measure of how much better a country might do. Neither the trend in record life expectancy nor the life expectancy trajectories in different countries suggest that a limit to life expectancy is in sight. Although rapid progress in catch-up periods is typically followed by slower increases, none of the curves appear to be approaching a maximum value (Oeppen and Vaupel 2002).

The rising numbers of centenarians in developed countries is another striking piece of evidence for the continuing increase in longevity. Lifespans exceeding 100 years, which seemed almost impossible to achieve in the past, despite spectacular reports, are increasingly becoming part of our reality today.

It is unlikely that any person living in Sweden before 1800 attained the age of 100 (Jeune 1995) and throughout the world centenarians must have been very rare (Wilmoth 1995). Data on the pre-18th century period have to be interpreted with caution. Few reliable statistics are available on mortality levels among the very old living under conditions of low life expectancy. The lower life expectancy is, the greater is the tendency to exaggerate age at older ages (Kannisto 1994). Today, the number of centenarians in developed countries is increasing at an exceptionally rapid rate of six to nine per cent per year in many countries. While 265 centenarians were counted in England and Wales in 1950, there were 5895 of them 50 years later, that is, more than 20 times the 1950 figure (Kannisto–Thatcher Database). In developed countries, the number of people celebrating their 100th birthday doubled each decade between 1950 and 1980; by the end of the 20th century it was multiplying by a factor of 2.4 per decade.

The History of Mortality Decline

How can the transition from high to low mortality be explained? Over most of the course of human existence, life expectancy hovered between 20 and 30 years. Infant mortality was high, people fell victim to infectious and parasitic diseases or simply to the harshness of everyday living conditions. Even in western Europe life expectancy did not reach age 40 until after 1800, and it stayed below age 50 until after 1900 (Vaupel and Jeune 1995). Over the course of the 20th century, life expectancy rose dramatically by more than 30 years in many industrialized countries. Rising life expectancy in industrialized countries since the 19th century is related to a fundamental epidemiological transition. There was a shift from the predominance of high mortality from infectious disease to conditions in which non-communicable and degenerative diseases among the elderly became more important. By the beginning of the

19th century in European areas of the world epidemics had been reduced, food supply became more stable, and fluctuations in mortality decreased. Over the course of the 19th century, the standard of living and hygiene improved and some public health services were established in a number of countries (Bongaarts and Bulatao 2000). Infectious disease was the greatest scourge of mankind until the first half of the 20th century, that is, until vaccination, antibiotics, and other medical advances finally began to combat successfully many of the life-threatening diseases in industrialized countries. By the same token, they lowered the rates of infant and child mortality and limited the devastating effects of the largest epidemics, although some outbreaks of influenza and the HIV/AIDS epidemic are exceptions. Parallel to these changes, there was a shift from high to low fertility. Mortality associated with pregnancy and birth decreased considerably.

The second half of the 20th century saw a dramatic reduction in death rates at advanced ages (Vaupel and Jeune 1995; Kannisto 1994; Kannisto et al. 1994; Vaupel 1997). The time around 1950 marks a distinct change in mortality conditions among the 'oldest old' (85 or more years of age) in developed countries: While improvements in survival were slow in the years preceding 1950, progress made after 1950 and especially after 1970 has been impressive. Data from England, Wales, France, Iceland, Japan, and the United States show clearly that old-age survival has been increasing since 1950 (Vaupel 1997; Vaupel et al. 1998). The population of centenarians and even supercentenarians (persons older than 110 years) is growing rapidly. The increase in the number of births about a century ago coupled with a sharp decline in mortality from childhood to age 80 contributed to the rising numbers. Demographic analyses, however, demonstrate that the most important factor behind the explosion of the centenarian population has been the decline in the mortality rate after age 80, a factor that has been two to three times more important than the other factors combined (Vaupel and Jeune 1995). The ongoing increase in life expectancy is largely attributable to continuous improvements in survival at advanced ages (Vaupel and Jeune 1995; Vaupel 1997).

In developed countries, the decline in mortality caused by infectious diseases and the postponement of degenerative diseases has delayed deaths to increasingly older ages. Today, cardiovascular disease and cancer are the major causes of death in industrialized countries. In 2002, heart disease and stroke accounted for more than half of all deaths, and cancers were responsible for around 20 per cent of all deaths in Europe (WHO 2004).

The human survival curve, which depicts the proportion of an initial (hypothetical) cohort still alive, has changed its shape as a consequence. The survival curve is becoming more rectangular due to the concentration of deaths at higher ages. To provide an example, the 2002 life table for Japanese women shows that more than 95 per cent of the initial hypothetical cohort would be still alive under current mortality rates at age 60. Mortality decline is neither a regular process in industrialized countries nor is it a process confined to these nations. Life expectancy has risen in most developing countries, too, especially in many Asian states and in Latin America. The mortality transition is driven by the same factors as in the developed countries – combating infectious disease plays a major role here. However, the transition proceeds much faster than it did in industrialized nations and there are considerable differences in the degree of progress (Bongaarts and Bulatao 2000).

The Plateau in Late-Life Mortality

Human death rates increase slowly after age 80. Data analyses of very large cohorts reveal that death rates reach a plateau at advanced ages and may level off around age 110 (Thatcher et al. 1998; Robine and Vaupel 2002). This observation is not unique to humans, however. Late-life mortality deceleration has been noticed in and confirmed for a number of model organisms as diverse as yeast, nematodes, or fruit flies. For all species for which large cohorts have been followed to extinction, age-specific mortality decelerates and, for the largest populations studied, even declines at older ages (Vaupel et al. 1998).

Some concepts contributing to an understanding of the astonishing improvement in survival at late ages come from biodemography, a subject that has emerged at the confluence of demography and biology. One biodemographic explanation builds on heterogeneity in frailty. All populations are heterogeneous, and even genetically identical populations display phenotypic differences. Frailer individuals have a lower probability of survival to late ages; robust individuals have a higher one. The frail tend to suffer high mortality, leaving a select subset of robust survivors. This results in compositional change in the surviving, aging population and in slower increases in age-specific death rates (Vaupel et al. 1979; Curtsinger et al. 1992; Vaupel and Carey 1993; Yashin et al. 1994). Another biodemographic explanation refers to changes in survival capacities at the individual level. Generally, the longevity of individual organisms is influenced by the living conditions to which they are exposed. Studies with different species have shown that several environmental factors of non-lethal stress, for example dietary restriction or heat shock, can induce increases in both resistance and longevity (Lithgow et al. 1995; Murakami and Johnson 1996; Masoro 2000). Hormesis, a biologically favourable response to low exposure to stress or toxins, is a well-known physiological phenomenon. Caloric restriction has proven to be an effective way to extend life span in a wide range of species, from yeast to mammals (Masoro 2000). It is not clear, however, whether fasting is a way of prolonging life in humans.

The Influence of Current Conditions on Age-Specific Death Rates

Studies involving model organisms have provided valuable insights into the biological processes of aging. An example can be drawn from a study on the Drosophila fruit fly. When flies fed a restricted diet were switched to a full diet, mortality soared to the level suffered by flies that had been fully fed all their lives. Conversely, when the diet of fully fed Drosophila was restricted, mortality plunged within 48 hours to the level enjoyed by flies that had experienced a lifelong restricted diet (Mair et al. 2003). The results support the repeated finding that age- specific death rates for humans (and other species) are strongly influenced by current conditions and behaviour (Kannisto 1994; Vaupel et al. 1998).

Placed in a broader context, the conclusion drawn from the fruit fly study also applies to humans. This can be illustrated neatly by an unplanned 'natural experiment' in Germany's recent history. Before reunification, both East and West Germany saw a radical decline in old-age mortality, as is characteristic for most developed countries. In the former GDR, however, mortality was considerably higher than in West Germany. Following unification (1989–1990), old-age mortality in East Germany declined to reach the levels prevailing in the West (Gjonca et al. 2000), a development largely attributed to improved health care for the elderly after unification. Thus, interventions even late in life can switch death rates to a lower, healthier trajectory. It's never too late to start prolonging your life (Vaupel et al. 2003).

Longevity in humans has a relatively low heritability. Studies of twins indicate that a modest 25 per cent of the variation in life spans is attributable to genetic differences among people (McGue et al. 1993; Herskind et al. 1996; Finch and Tanzi 1997). The discoveries of genetic and environmental factors that contribute to extensions of the lifespan do not fully explain the malleability of aging. Nevertheless, the findings show that there are means and ways of delaying aging.

The Plasticity of Aging

The rise in life expectancy has provoked discussion of the question whether we are approaching a limit to life expectancy, a biologically determined maximum lifespan that inevitably halts further improvements of old-age survival.

A common assumption still widely held is that lifespan cannot be extended beyond a biologically determined limit. The notion of an inevitable maximum lifespan also influences scientific studies of longevity (Fries 1980; Olshansky et al. 1990).

Ever since research into longevity began, attempts have been made to determine the maximum life expectancy that humans could reach. The ceilings proposed by various authors differ but all have been exceeded, apart from those proposed most recently (Oeppen and Vaupel 2002). The assumption of a finite, biological limit to life can be traced back to Aristotle (350 BC). In his treatise 'On Youth and Old Age, On Life and Death', Aristotle contrasted two types of death: premature death caused by disease or accident, and senescent death due to old age. He believed that nothing could be done about old age and thus about the end to life. More than 2300 years later, James Fries quantified Aristotle's distinction in a widely cited article published in the *New England Journal of Medicine.* If life is not cut short by accident or illness, then the lifespan of man will inevitably approach a potential maximum limit that is fixed for every human but differs from individual to individual (Fries 1980). According to Fries, the fixed value of the maximum lifespan is normally distributed with a mean of 85 years and a standard deviation of seven years. Fries emphasizes that nothing can be done to alter a person's maximum lifespan as the latter is beyond the influence of environmental, behavioural, or medical intervention currently conceivable. Accordingly, death rates at older ages are intractable. The notion of unavoidable senescent death has been reinforced by evolutionary biologists who hypothesize that mortality must rise with age as the force of selection against deleterious, late-acting mutations declines (Hamilton 1966).

The notion of an upper biological limit to lifespan may be commonly accepted, yet there is no empirical evidence of a proximate limit to human longevity. The steady rise in human life expectancy shows no signs of levelling off. Experts repeatedly asserting that life expectancy is approaching a ceiling have repeatedly been proven wrong. If life expectancy were approaching an unavoidable biologically maximum, then the increase in life expectancy should be slowing, especially in countries such as Japan or France, both of which enjoy exceptionally low death rates. This, however, is not the case (Oeppen and Vaupel 2002; Vaupel 1997). Mortality is plastic even at advanced ages.

The prevailing causes of rising life expectancy have undergone changes and are complex. Combined, they have nonetheless led to a stable and linear increase in life expectancy since 1840. This will probably also apply to the future. Just as medical breakthroughs – for example, the discovery of antibiotics or advances in organ transplantation – were not foreseen, we do not know what major technological innovations the future will bring to promote long and healthy lives. There is no reason, however, to assume that progress in technological knowledge and its exploitation will come to a halt. It would not make sense to take the standards of today to estimate the conditions influencing life expectancy tomorrow. Future advances in life expectancy will be made as we progress in the prevention, diagnosis, and treatment of deadly age-related diseases (Barbi and Vaupel 2005).

Future Prospects of Longevity

Because best-practice life expectancy has been increasing by 2.5 years per decade for the past 160 years, one reasonable scenario is that this trend will continue in the coming decades. To date, there is no indication that a change in the trend is in sight. If the trend continues, there may be a country in about six decades' time with life expectancy beyond the threshold of 100 years (Oeppen and Vaupel 2002).

An application of this extrapolation in conjunction with methods from time- series analysis to project the gap between best-practice and national life expectancy results in national forecasts that are considerably higher than many official projections. From the use of this method, female life expectancy for Germany, for example, is expected to rise significantly above 90 years by 2050. Official projections, however, do not exceed 87 years (medium scenario). In many countries, official projections assume a deceleration in reductions of death rates. Such projections made in the past have resulted in underestimates of actual increases in life expectancy. These errors

distort planning for future pensions, health care, and other social needs as well as the decision-making of individuals drawing up saving plans or planning for retirement. Increases in life expectancy of a few years can produce large changes in the numbers of old and oldest old who will need support and care. In developed countries, centenarians may well become commonplace during the lifetime of people alive today.

Mortality Divergences

Although health trends have been generally positive throughout the world and remarkable improvements in survival have been achieved in developed and many developing countries (Tuljapurkar et al. 2000; Vallin and Meslé 2005), death rates still vary among countries and even within countries. In the 1970s and early 1980s, many demographers expected a convergence in life expectancies worldwide by assuming gains would be higher for the countries with lower life expectancies (McMichael et al. 2004). A quarter of a century later, however, it is clear that this assumption did not hold. On the one hand, increases in life expectancy of some of the best-performing countries, such as Japan or France, did not show any levelling off at all and life expectancy climbed higher than expected. On the other hand, there have been exceptions to the widespread phenomenon of general mortality decline in the second half of the 20th century. Mortality reversals were observed in the 1980s and 1990s in as many as 42 countries (McMichael et al. 2004; Caselli et al. 2002; Vallin and Meslé 2005) as life expectancy fell. Most of these countries are situated in sub-Saharan Africa or in eastern Europe. Life expectancy in several sub-Saharan countries was more than ten years lower in 2004 than predicted by the UN Population Division about 20 years earlier (United Nations 1981). Other countries that experienced reversals in life expectancy at the end of the 20th century are North Korea, Haiti, Fiji, the Bahamas, and Iraq. Setbacks apart from those caused by war and famine were not taken into account by early demographers, with the result that future setbacks in national mortality were considered unlikely (McMichael et al. 2004).

In sub-Saharan Africa, HIV/AIDS and other infectious diseases, such as tuberculosis and malaria, caused death rates to rise, and many of the countries involved were additionally faced with economic hardships, political conflicts, and violence between groups or individuals. Russia, like other countries of the former USSR or of eastern Europe, experienced increased mortality among working-age adults, especially among men aged between 20 and 65 (Shkolnikov et al. 1998; Meslé et al. 2003). Adults are normally less vulnerable to mortality increase than are children or the elderly. The drastic political and socio-economic transition increased unemployment rates and income inequalities, and led to weakened safety nets and to psycho-social stress among those most affected, particularly the less educated population groups (Shapiro 1995; Shkolnikov et al. 1998; Bobak et al. 2000). Adverse male behaviours, such as alcohol abuse, crime, and violence, contributed to male excess adult mortality. In addition, rates of cardiovascular disease and cancer mortality are high in Russia.

Some industrialized countries perform less well than others. Since the mid- 1980s in the United States, for example, death rates have declined more slowly than in most other developed countries. Until about 1980, the United States enjoyed relatively low death rates for both women and men after aged 65. Since then, however, death rates at older ages have fallen less rapidly than in Japan, France and other countries. The reasons for the slow increase in life expectancy in the United States are not yet well understood.

Mortality Differentials

The U-shape of the mortality risk trajectory applies to all humans. Nevertheless, remarkable differentials exist by geographical region and along other dimensions. The best-known differential is between females and males. In most developed countries, the difference between female and male life expectancy is between four and seven years. The gap between women and men is

typically smaller in less developed countries. It is not clear how much of the gap is biological as opposed to social, in part because biological factors interact with social ones. While men take more health risks (such as smoking), women are more careful about their health (for example, visits to the doctor).

Socio-economic status (SES) and mortality have an inverse relationship: individuals with higher SES usually enjoy lower mortality, regardless of how SES is measured (Goldman 2001). Although measures of SES are correlated with each other, they address different dimensions: education is related to health behaviour and knowledge of healthy lifestyle, occupation to health hazards of the job, and income to access to health care as well as to the ability to provide a healthy living environment (such as housing conditions).

Marital status is another important mortality determinant. Married individuals usually have lower death rates than do never-married women and men, the widowed, or the divorced. Two different hypotheses have been discussed in the literature to explain this differential. On the one hand, marriage is expected to have a protective effect via pooled financial resources, higher social support, the adoption of healthier lifestyles, and other factors. On the other hand, it is argued that there is a selection effect into marriage: healthy women and men have higher chances of finding a spouse than less healthy individuals (Goldman 1993).

See Also

- ▶ Fertility in Developed Countries
- ▶ Fertility in Developing Countries
- ▶ Retirement

Bibliography

Aristotle. 350 BC. *On youth and old age, on life and death, on breathing*. Trans. G. Ross. http://classics.mit.edu/Aristotle/youth_old.html. Accessed 20 Aug 2006.

Barbi, E., and J. Vaupel. 2005. Comment on 'Inflammatory exposure and historical change in human life-spans'. *Science* 308: 1743.

Bobak, M., H. Pikhart, C. Hertzman, R. Rose, and M. Marmot. 2000. Socioeconomic factors, material inequalities, and perceived control in self-rate health: Cross-sectional data from seven post-communist countries. *Social Science and Medicine* 51: 1343–1350.

Bongaarts, J., and R. Bulatao, eds. 2000. *Beyond the billion: Forecasting the world's population*. Washington, DC: National Academy Press.

Caselli, G., F. Meslé, and J. Vallin. 2002. Epidemiologic transition theory exceptions. *Genus* 58 (1): 9–51.

Curtsinger, J., H. Fukui, D. Townsend, and J. Vaupel. 1992. Demography of genotypes: Failure of the limited life-span paradigm in *Drosophila melanogaster*. *Science* 258: 461–463.

Finch, C., and R. Tanzi. 1997. Genetics of aging. *Science* 278: 407–411.

Fries, J. 1980. Aging, natural death, and the compression of morbidity. *New England Journal of Medicine* 303: 130–135.

Gjonca, A., H. Brockmann, and H. Maier. 2000. Old-Age mortality in Germany prior to and after reunification. *Demographic Research* 3(1). http://www.demographic-research.org/volumes/vol3/1/3-1.pdf. Accessed 20 Aug 2006.

Goldman, N. 1993. Marriage selection and mortality patterns: Inferences and fallacies. *Demography* 30: 189–208.

Goldman, N. 2001. Mortality differentials: Selection and causation. In *International encyclopedia of the social and behavioral sciences*, ed. N. Smelser and P. Baltes. Oxford: Elsevier Science.

Gompertz, B. 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London* 11: 513–583.

Hamilton, W. 1966. The moulding of senescence by natural selection. *Journal of Theoretical Biology* 12: 12–45.

Heligman, M., and J. Pollard. 1980. The age pattern of mortality. *Journal of the Institute of Actuaries* 107: 49–80.

Herskind, A., M. McGue, T. Soerensen, and J. Vaupel. 1996. The heritability of human longevity: A population-based study of 2872 Danish twin pairs born 1870–1900. *Human Genetics* 97: 319–323.

Jeune, B. 1995. In search for the first centenarians. In *Exceptional longevity: From prehistory to the present*, ed. B. Jeune and J. Vaupel. Odense: Odense University Press.

Kannisto, V. 1994. *Development of the oldest-old mortality, 1950–1990*. Odense: Odense University Press.

Kannisto, V., J. Lauritsen, A. Thatcher, and J. Vaupel. 1994. Reductions in mortality at advanced ages: Several decades of evidence from 27 countries. *Population and Development Review* 20: 793–730.

Kannisto–Thatcher Database on old age mortality. http://www.demogr.mpg.de/databases/ktdb/. Accessed 20 Aug 2006.

Lithgow, G., T. White, S. Melov, and T. Johnson. 1995. Thermotolerance and extended life-span conferred by single-gene mutations and induced by thermal stress.

Proceedings of the National Academy of Sciences USA 92: 7540–7544.
Mair, W., P. Goymer, S. Pletcher, and L. Partridge. 2003. Demography of dietary restriction and death in *Drosophila*. *Science* 301: 1731–1733.
Masoro, E.J. 2000. Caloric restriction and aging: An update. *Experimental Gerontology* 35: 299–305.
McGue, M., J. Vaupel, N. Holm, and B. Harvald. 1993. Longevity is moderately heritable in a sample of Danish twins born 1870–1880. *Journals of Gerontology: Series A, Biological Sciences and Medical Sciences* 48: B237–B244.
McMichael, A., M. McKee, V. Shkolnikov, and T. Valkonen. 2004. Mortality trends and setbacks: Global convergence or divergence? *The Lancet* 262: 1155–1159.
Meslé, F., J. Vallin, V. Hertrich, E. Andreev, and V. Shkolnikov. 2003. Causes of death in Russia: Assessing trends since the 1950s. In *Population of Central and Eastern Europe. Challenges and opportunities*, ed. I. Kotowska and J. Joswiak. Warsaw: Statistical Publishing Establishment.
Murakami, S., and T. Johnson. 1996. A genetic pathway conferring life extension and resistance to UV stress in *Caenorhabditis elegans*. *Genetics* 143: 1207–1218.
Oeppen, J., and J. Vaupel. 2002. Broken limits to life expectancy. *Science* 296: 1029–1031.
Olshansky, S., B. Carnes, and C. Cassel. 1990. In search of Methuselah: Estimating the upper limits of human longevity. *Science* 250: 634–640.
Riley, J. 2001. *Rising life expectancy: A global history*. Cambridge: Cambridge University Press.
Robine, J., and J. Vaupel. 2002. Emergence of supercentenarians in low mortality countries. *North American Actuarial Journal* 6: 54–63.
Shapiro, J. 1995. The Russian mortality crisis and its causes. In *Russian economic reform at risk*, ed. A. Aslund. London: Pinter.
Shkolnikov, V., G. Cornia, D. Leon, and F. Meslé. 1998. Causes of the Russian mortality crisis: Evidence and interpretations. *World Development* 26: 1995–2011.
Thatcher, A., V. Kannisto, and J. Vaupel. 1998. *The trajectory of mortality from age 80 to 120*. Odense: Odense University Press.
Tuljapurkar, S., N. Li, and C. Boe. 2000. A universal pattern of mortality decline in the G7 countries. *Nature* 405: 789–792.
United Nations. 1981. *World population prospects as assessed in 1980*. New York: United Nations.
United Nations. 2005. *World population prospects, the 2004 revision, highlights*. New York: Department of Economic and Social Affairs, United Nations.
Vallin, J., and F. Meslé. 2005. Convergences and divergences: An analytical framework of national and subnational trends in life expectancy. *Genus* 61 (1): 83–123.
Vaupel, J. 1997. The remarkable improvements in survival at old ages. *Philosophical Transactions of the Royal Society of London, Series B* 352: 1799–1804.
Vaupel, J., and J. Carey. 1993. Compositional interpretations of medfly mortality. *Science* 260: 1666–1667.
Vaupel, J., and B. Jeune. 1995. The emergence and proliferation of centenarians. In *Exceptional longevity: From prehistory to the present*, ed. B. Jeune and J. Vaupel. Odense: Odense University Press.
Vaupel, J.W., K.G. Manton, and E. Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16: 439–454.
Vaupel, J., J. Carey, K. Christensen, T. Johnson, A. Yashin, N. Holm, I. Iachine, V. Kannisto, A. Khazaeli, P. Liedo, V. Longo, Y. Zeng, K. Manton, and J. Curtsinger. 1998. Biodemographic trajectories of longevity. *Science* 280: 855–860.
Vaupel, J.W., J. Carey, and K. Christensen. 2003. It's never too late. *Science* 301: 1679–1681.
WHO. 2006. *World health report 2006: Working together for health*. Geneva: WHO.
WHO (World Health Organization). 2004. *Maternal mortality in 2000: Estimates developed by WHO, UNICEF, and UNFPA*. Geneva: WHO.
Wilmoth, J. 1995. The earliest centenarians: A statistical analysis. In *Exceptional longevity: From prehistory to the present*, ed. B. Jeune and J. Vaupel. Odense: Odense University Press.
Yashin, A., J. Vaupel, and I. Iachine. 1994. A duality in aging: The equivalence of mortality models based on radically different concepts. *Mechanisms of Ageing and Development* 74: 1–14.

Free Internet Sources on Mortality Data

The Human Life-Table Database is a collection of life tables for more than 30 countries. http://www.lifetable.de. Accessed 20 Aug 2006.
The Human Mortality Database contains calculations of the death rates and life tables of almost 30 countries. Access to the data requires registration, which is free at http://www.mortality.org. Accessed 20 Aug 2006.
The Kannisto-Thatcher Database on Old Age Mortality is tailored to analyse mortality at ages 80 and over. It has death and population counts by sex, age, birth year, and calendar year for more than 30 countries for ages 80 and above. http://www.demogr.mpg.de/databases/ktdb. Accessed 20 Aug 2006.

Mortensen, Dale T. (Born 1939)

Rasmus Lentz

Abstract

Dale T. Mortensen (born 1939) was awarded the Nobel Prize for Economics in 2010 jointly

with Peter A. Diamond and Christopher A. Pissarides for his work on the analysis of markets with search frictions. Together, they developed the Diamond-Mortensen-Pissarides Model (DMP model): an equilibrium model of unemployment dynamics. An empirically motivated theoretical economist, Mortensen has made a tremendous contribution to the field of labour economics.

Keywords

Nobel Prize; Wage search model; Market tightness; Worker reallocation; Labour market theory; Labour market search; Diamond-Mortensen-Pissarides model; Beveridge curve

JEL Classifications

B31

Introduction

A product of the Pacific Northwest and Scandinavian heritage, Dale T. Mortensen was born in 1939 in Enterprise, Oregon, the first of three sons of Verna Ecklund and Thomas Peter Mortensen. In 2010, together with Peter A. Diamond and Christopher A. Pissarides, Mortensen was awarded the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for their analysis of markets with frictions. In 1961 Mortensen entered Carnegie Mellon University to pursue graduate studies in economics. It was in Pittsburgh that he met his wife, Beverly Patton. They were married in 1963 and in quick succession had three children, Karl, Lia and Julie. Mortensen joined the Northwestern faculty in 1965 and finished his PhD thesis a couple of years later under the primary supervision of Michael Lovell and Alan Meltzer. Mortensen's unions with Beverly and Northwestern remain to this day. Over the course of his career, Mortensen has taught at Essex University, Cornell University, the California Institute of Technology and New York University. From 2005 to the present, Mortensen has also been affiliated with Aarhus University as the Niels Bohr Visiting Professor of Economics.

Early Developments of Search Theory: The Wage Search Model

Mortensen's pioneering work on the outcomes of a decentralized, frictional meeting process between firms and workers is reflected in two papers, both published in 1970: 'A theory of wage and employment dynamics', is part of the famous 'Phelps Volume' whose contributors by now have three Nobel Prizes between them, and 'Job search, the duration of unemployment and the Phillips curve,' published in the *American Economic Review*. The work was motivated by the vigorous debate in the mid- and late 1960s about the relationship between unemployment and inflation as embodied in the Phillips curve. From this perspective, Mortensen's roots in the truly remarkable Carnegie Mellon group of the mid-1960s show very clearly his attempt to build a microeconomic foundation for macroeconomic policy and analysis. However, these groundbreaking articles are not primarily remembered for their contributions to the Phillips curve debate. Rather, together with John McCall's (1970) article in the *Quarterly Journal of Economics*, they lay the foundation for the wage search model, and with it frictional unemployment theory. The

next section on the Diamond, Mortensen and Pissarides (DMP) model on search and matching provides a more detailed description of frictional unemployment theory. The wage search model described in this section is the cornerstone of frictional foundations in the DMP model.

In its simplest form, the wage search model describes the unemployed worker's process of finding an acceptable job. The unemployed worker faces a market in which employment opportunities differ in the wages they offer. The worker must go through a search process to locate opportunities and not until actual inspection will the worker know how attractive each opportunity is. Search is modelled as a process of sequential sampling where opportunities arrive stochastically according to a Poisson process, i.e. it is assumed that job opportunities arrive continuously and independently of each other. The Poisson arrival rate, denoted by λ, dictates the frequency by which offers arrive. The passing of time between offer arrivals defines friction in the model.

As a side note, in today's modern labour market theory, researchers give little thought to the Poisson offer arrival process method of modelling search frictions. It has become about as fundamental as breathing. However, it is a crucial modelling choice as it allows simple aggregation of an advanced micro foundation of frictional job search into aggregate labour market dynamics. Mortensen's work is credited with the introduction of the Poisson arrival process as a way of modelling search frictions, and it is a major part of why we are today working with relatively simple macro models of frictional labour markets that have solid microeconomic foundations.

Returning to the wage search model, an employment opportunity which is fully characterized by its wage rate, w, is a sample from the cumulative wage offer distribution in the market, $F(w)$. For expositional purposes only, assume time is continuous, jobs last forever, that individuals live infinitely long with a discount rate of r, and that individuals are income maximizers. The asset value of a job with wage w is then $W(w) = w/r$. During unemployment a worker receives income at rate b and if she searches, she pays search cost at rate c. The asset value of unemployment is,

$$rU = b - c + \lambda \int [\max[W(w), U] - U] dF(w), \quad (1)$$

which says that the dividend flow of being an unemployed worker engaged in search is the instantaneous income flow $b - c$ plus the dividend flow from the job opportunity arrival process. If the value of an employment opportunity exceeds unemployment, $W(w) \geq U$ then the unemployed worker accepts the offer. Otherwise the offer is rejected and the search continues. The acceptance/rejection decision is the key behavioural implication of the model.

The value of employment is monotonically increasing with wage. Consequently, the acceptance/rejection decision reduces to a simple threshold decision: Accept all wage offers $w \geq R$, and reject all offers below the threshold, also referred to as the reservation wage. The reservation wage is defined uniquely by $W(R) = U$. In general, the reservation threshold strictly exceeds the unemployed income flow, b, because employment implies the loss of the option to search for outside employment opportunities. With these insights, Eq. 1 can be rewritten to reflect the reservation level,

$$R = b - c + \frac{\lambda}{r} \int_R [w - R] dF(w). \quad (2)$$

Equation 2 allows straightforward determination of comparative statics on the reservation wage: the reservation wage increases with the income flow net of search costs, $b - c$, increases with offer arrival rate, λ, increases with mean preserving spreads and right translations of the offer distribution, and decreases with the discount rate.

To this day, the model and its refinements are the theoretical foundation of the study of unemployment durations. Specifically, the worker leaves unemployment at rate $\lambda[1 - F(R)]$.

The average unemployment duration is the inverse of this rate.

The early wage search model had prominent critics. James Tobin argued that an important part of hiring is directly from one job to another. The impact of this criticism is profound in that if search while in a job is just as efficient as unemployed search, then the acceptance/rejection decision trivializes to the acceptance of any wage offer in excess of the unemployed income flow, *b*. The implication of which is that the acceptance/rejection decision is irrelevant to the understanding of unemployment durations. This resulted in the rise of the other key behavioural prediction of the search model, the search intensity decision: that the worker can engage in costly search to affect the rate at which employment opportunities arrive. Ken Burdett developed the first formal model of on-the-job search in his PhD thesis, published as Burdett (1978). Not only is the search intensity decision an important determinant of unemployment durations, Christensen et al. (2005) showed that it is an important part of the mechanism that reallocates workers between jobs.

A more fundamental and lasting question that was raised concerns the source of wage dispersion in the search model. Michael Rothschild posed the question in Rothschild (1973) and the question is at the core of the motivation in Diamond (1971), where Peter Diamond determined that the equilibrium wage distribution in a simple sequential search model is degenerate at the monopsony wage. Burdett and Judd (1983) provided a classic answer to the question, and shortly thereafter Burdett 1998 provided the natural extension to on the job search. It took the better part of a decade for Burdett 1998 to find its place at the *International Economic Review*. Equilibrium wage dispersion in markets with frictions is a lasting topic in Mortensen's work. In 2000, Mortensen gave the Zeuthen Lectures at University of Copenhagen on the topic, which were later published in the excellent Mortensen (2003).

Search and Matching: The DMP Model

In the early 1980s, Mortensen explored two-sided search and the notion of a matching function where the aggregate number of matches in the market is 'produced' from matching efforts both on the side of unemployed workers and on the firm side through vacancies. Peter Diamond and Christopher Pissarides were engaged in similar efforts during this time as well. The work is contained in Diamond (1982a, b), Mortensen 1982a, b), and Pissarides (1985). Christopher Pissarides in particular is credited with adding the free entry condition in vacancy creation to the model and thereby endogenizing the firm side of match creation. The early work on the model does not explicitly deal with the job destruction decision. Job destruction in the DMP model is studied in Mortensen and Pissarides (1994), which also reflects Mortensen's steady insistence on taking inspiration from labour market data and creating models that speak directly to the data. In this case, the model is related to empirical work on job creation and destruction as described in Davis et al. (1996).

The Diamond-Mortensen-Pissarides model is an analysis of frictional unemployment dynamics. Pissarides (2000) provides an excellent treatment of the model. Frictional unemployment theory is sometimes referred to as an equilibrium theory of unemployment with the often stated contrast of Keynesian unemployment theory, where unemployment is a result of market wages not adjusting downwards to equate labour supply with labour demand. In frictional unemployment theory, unemployment is a manifestation of the time-consuming process of workers and employers finding each other to form matches. Unemployment moves to the extent that changes in the economy impose themselves on resources dedicated to job creation by firms and job search by workers, or if the frictional process itself changes. Wages are set by bilateral bargaining and market conditions affect wages through their impact on match surplus and outside options. The model describes behaviour at the individual level and unemployment dynamics are the result of the aggregation of individual behaviour to the economy-wide level. The model is a revolution to the study of unemployment dynamics much the same way that macroeconomics was modernized by the introduction of micro foundations. It is in

this sense that one can reasonably trace kinship back to the first Phelps volume.

At the core of the DMP model is the matching function, $m = m(u,v)$, where m is the matching rate, u is the unemployment rate and v is the vacancy rate, all relative to a fixed labour force, L. The typical analysis will impose basic regularity conditions on the matching function: (1) The matching function is increasing in each of its arguments, (2) The matching function is homogeneous of degree one. Market tightness is defined by $\theta = v/u$. A vacancy is matched with an unemployed worker at rate $q(\theta) = m(u,v)/v = m(1/\theta,1)$. The vacancy matching rate is decreasing in market tightness. An unemployed worker is matched with a vacancy at the job finding rate $f(\theta) = m/u = \theta q(\theta)$, which is increasing in market tightness. In the simple version of the model, jobs are destroyed exogenously at rate δ. Upon job destruction, the worker is laid off into the unemployment pool. For expositional purposes, assume a simple two-state economy where workers can be either employed or unemployed. There is no labour force participation decision.

Firms are one match units. A match between a worker and a firm produces revenue flow p. An unemployed worker has income flow b, net of search costs. A vacancy costs pc at any point in time, where $c > 0$. Once a match is filled, the worker receives income w at any point in time and the firm takes the residual, $p - w$. With income maximizing agents, a discount rate of r and in continuous time, the worker's asset values of being unemployed (U) and employed (W) are,

$$rU = b + f(\theta)[W - U]$$

$$rW = w - \delta[W - U].$$

The firm's asset values of a vacancy (V) and a job (J) are,

$$rV = -pc + q(\theta)[J - V]$$

$$rJ = p - w - \delta[J - V].$$

Market tightness and wages are determined by assumption of free entry of vacancies and through Nash bargaining. Free entry in vacancy creation implies that at any point in time vacancies will be added to or withdrawn from the stock of vacancies to the point where

$$V = 0. \tag{3}$$

The firm and the worker split match surplus according to generalized Nash bargaining. Define the match surplus by $S + W + J - U$, where the free entry condition in Eq. 3 has been applied. Consequently, the wage level is at any point set so that

$$W = U + \beta S, \tag{4}$$

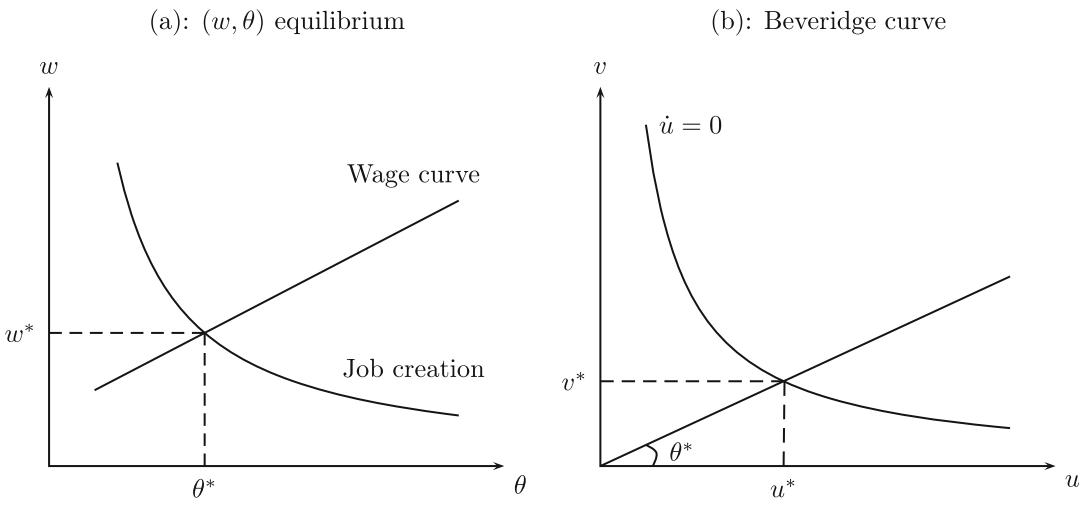
where β is the worker's bargaining power.

Equilibrium is defined as a combination of the wage and market tightness (w, θ) that satisfies the free entry and Nash bargaining conditions in Eqs. 3 and 4 for the given asset value definitions. The equilibrium is often illustrated as the intersection between two curves in (w,θ) space: the job creation curve and the wage curve. The job creation curve follows directly from the free entry condition in Eq. 3 and the definitions of V and J,

$$p - w = \frac{(r + \delta)pc}{q(\theta)}. \tag{5}$$

The job creation curve describes a negative relationship between the wage and market tightness. A lower wage increases the firm's valuation of a filled vacancy. Consequently, the value of a vacancy is greater, inducing additional vacancy creation and thereby an increased market tightness.

The wage curve is obtained in two steps. First, by the asset value definitions, the wage bargaining equation can be stated in its flow equivalent, $w = rU + \beta(p - rU)$. The second step relates the worker's valuation of unemployment to market tightness by combining both the free entry condition and the wage bargaining equation with the asset value definitions to obtain, $rU = b + \beta pc\theta/(1 - \beta)$. Together, the two equations yield the wage equation



Mortensen, Dale T. (Born 1939), Fig. 1 Steady state equilibrium in the DMP model.

$$w = (1 - \beta)b + \beta p(1 + c\theta). \tag{6}$$

The wage equation describes a positive relationship between wages and market tightness. In a tighter market the worker's job finding rate increases, thereby increasing the value of unemployment and with it the worker's wage bargaining position. The equilibrium is described in Fig. 1a.

Unemployment dynamics in the DMP model are described by the differential equation

$$\dot{u} = (1 - u)\delta - uf(\theta), \tag{7}$$

which simply states that the unemployment rate changes according to the rate at which workers flow into unemployment from employment minus the outflow rate. Steady state is defined as the state where unemployment does not change over time, hence $\dot{u} = 0$. Impose steady state on Eq. 7 and one obtains the Beveridge curve:

$$u = \frac{\delta}{\delta + \theta q(\theta)}, \tag{8}$$

which is unemployment–vacancy combinations that satisfy steady state. The Beveridge curve is illustrated in Fig. 1b.

Steady state equilibrium in the DMP model is a combination of the equilibrium conditions on (w,θ) and the Beveridge curve in Eq. 8, as illustrated in Fig. 1. For the purposes of business cycle analysis, the comparative static in p is of particular interest as a measure of a typical supply shock. An increase in p shifts the wage curve up and the job creation curve to the right. The job creation curve shifts by more so that both wages and market tightness increase. This then results in a move to the north-west along the Beveridge curve, resulting in lower unemployment and a greater measure of vacancies. Mortensen and Pissarides (1994) analyze endogenous job destruction along with an aggregate shock process that allows description of transitional dynamics over the business cycle. It is a key point that with endogenous job destruction, the Beveridge curve relationship between unemployment and vacancies is not stable over the business cycle. Supply shocks not only induce movement along the Beveridge curve, but also a shift of the curve because the job destruction rate changes.

During the 1990s Mortensen and Chris Pissarides produced a series of highly successful papers that implement the DMP model into the study of unemployment dynamics over the business cycle as well as the study of a wide range of government policies. As a result, the model gradually became a standard workhorse in the macroeconomic business cycle and policy analysis whenever a serious labour market description

was required. A fact that is underscored by the great amount of attention given to Robert Shimer's (2005) paper on wage and unemployment volatility. Here, Shimer documents that the DMP model in its basic form cannot reproduce the observed covariance magnitude between wages, vacancies and unemployment over the business cycle. Shimer suggests that the likely culprit is the assumed wage determination mechanism in the model which implies an overly sensitive wage response to labour market tightness changes over the business cycle. In work with Eva Nagypal, Mortensen suggested a number of alternatives focused on producing relatively small firm profit rates which may as a result be more elastic to supply shock innovations. On-the-job search is one such mechanism. Overall, Shimer's paper has stimulated papers too numerous to cite here aimed at modifying the DMP model to bring it in line with the empirical evidence on wage and unemployment volatility.

Match Data, Worker Reallocation and Labour Market Heterogeneity

Over the past one or two decades, Mortensen's research has increasingly focused on building a framework to understand and quantify the importance of worker reallocation. As has been repeatedly documented, at any point in time, regardless of the state of the economy, the labour market is doing a tremendous amount of costly reallocation of workers between jobs with or without intervening unemployment spells. Papers like Christensen et al. (2005) and Lentz and Mortensen (2008) argue that worker reallocation is an efficient response to friction and the changing fortunes of firms. As argued above, early work on labour market friction as well as the DMP model is motivated by the desire to understand unemployment in relation to other aggregate economic measures. Of course, a substantial part of the evaluation of the performance of a given labour market concerns its ability to match unemployed workers with jobs. However, quite possibly just as important a concern is its ability to facilitate the ongoing reallocation of workers away from firms in decline to more innovative and productive firms. To the extent that there is such a thing as the 'right job' for each worker, labour market performance is measured in part by how well it facilitates the worker's quest for that job.

Labour economics is a field with a substantial interaction between theoretical and empirical development. Mortensen is a particularly strong example of an empirically motivated theoretical economist. As a result, over the years he has gathered around him an exceptionally broad group of researchers from the purely empirical to the purely theoretical, all of whom have benefited tremendously from his input.

Mortensen's empirical work is primarily based on Danish data. The connection to Danish empirical research groups goes back to the early 1980s. In the late 1970s and early 1980s Mortensen worked together with Ken Burdett, Nick Kiefer and George Neumann on interpreting worker history data through search models. Specifically, the research focus is on spell duration analysis and wage offer distribution estimation. Lars Muus, a PhD student in a Nordic PhD course that Mortensen and Burdett teach in Oslo in 1980 informed them of a Danish project headed by Henning Bunzel and Niels Westergaard-Nielsen on worker history data collected from the Danish Union of Jurists and Economists. They subsequently invite Burdett, Kiefer, Neumann and Mortensen to a conference on Danish worker history data at Sandbjerg Manor in 1982. This marks the beginning of an ongoing collaboration between Mortensen and his Danish colleagues on the development of micro panel data for labour market studies. In the late 1990s the data developed into what is now referred to as matched employer–employee (MEE) data based on administrative records. The core observation in matched employer–employee data is a match between a worker ID and an employer ID along with various match characteristics. This is typically linked with worker and employer characteristics. MEE data sets are now available for a broad set of countries.

The study of worker reallocation has taken Mortensen and Lentz in the direction of identifying heterogeneity on the demand side of the labour market in particular, but possibly also on the supply side, as a core condition for the importance of worker reallocation in the presence of labour market frictions and/or the firm growth dynamics of an

innovative economy. It has motivated us to shine a stronger light on the decision process of the firm than is typical in labour economics in order to understand the dynamics of labour demand. In some ways, the research agenda brings us full circle to Mortensen's early work on the dynamics of the firm's factor demands in the presence of adjustment costs, Mortensen (1973). This happens to be Mortensen's first *Econometrica* publication. Thirty-five years later he would return to the study of firm dynamics in the same journal in Lentz and Mortensen (2008).

See Also

- ▶ Labour Market Search
- ▶ Layoffs
- ▶ Matching
- ▶ Pissarides, Christopher (Born 1948)
- ▶ Search Models of Unemployment
- ▶ Search Theory

Bibliography

Burdett, K. 1978. A theory of employee job search and quit rates. *American Economic Review* 68(1): 212–220.

Burdett, K., and K.L. Judd. 1983. Equilibrium price dispersion. *Econometrica* 51(4): 955–969.

Burdett, K., and D.T. Mortensen. 1998. Wage differentials, employer size, and unemployment. *International Economic Review* 39(2): 257–273.

Christensen, B.J., R. Lentz, D.T. Mortensen, G. Neumann, and A. Werwatz. 2005. On the job search and the wage distribution. *Journal of Labor Economics* 23(1): 31–58.

Davis, S.J., J.C. Haltiwanger, and S. Schuh. 1996. *Job creation and destruction*. Cambridge/London: MIT Press.

Diamond, P.A. 1971. A model of price adjustment. *Journal of Political Economy* 3(2): 156–168.

Diamond, P.A. 1982a. Aggregate demand management in search equilibrium. *The Journal of Political Economy* 90(5): 881–894.

Diamond, P.A. 1982b. Wage determination and efficiency in search equilibrium. *Review of Economic Studies* 49(2): 217–227.

Lentz, R., and D.T. Mortensen. 2008. An empirical model of growth through product innovation. *Econometrica* 76(6): 1317–1373.

McCall, J.J. 1970. Economics of information and job search. *Quarterly Journal of Economics* 84(1): 113–126.

Mortensen, D.T. 1970a. Job search, the duration of unemployment and the Phillips curve. *American Economic Review* 60(5): 847–862.

Mortensen, D.T. 1970b. A theory of wage and employment dynamics. In *Microeconomic foundations of employment and inflation theory*, ed. E.S. Phelps. New York: W. W. Norton and Co.

Mortensen, D.T. 1973. Generalized costs of adjustment and dynamic factor demand theory. *Econometrica* 41(4): 657–665.

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

Mortensen, D.T. 1982b. Property rights and efficiency in mating, racing, and related games. *American Economic Review* 72(5): 968–979.

Mortensen, D.T. 2003. *Wage dispersion: Why are similar workers paid differently?* Cambridge/London: MIT Press.

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

Pissarides, C.A. 1985. Short-run equilibrium dynamics of unemployment, vacancies and real wages. *American Economic Review* 75(4): 676–690.

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

Rothschild, M.D. 1973. Models of market organization with imperfect information: A survey. *Journal of Political Economy* 81(6): 1283–1308.

Shimer, R. 2005. The cyclical behavior of equilibrium unemployment and vacancies. *American Economic Review* 95(1): 25–49.

Motion Pictures, Economics of

W. David Walls

Abstract

Film-goers discover the films they like by consuming them, and through the exchange of information the demand for motion pictures evolves dynamically. The supply of screens adjusts in response to demand through flexible state-contingent exhibition contracts. This article presents an overview of the economics of motion pictures that focuses on how the demand process affects the distribution of outcomes, how the distribution of outcomes can

be quantified with the use of statistical models, and how the industry's organization and business practices can be understood in light of the behavioural and statistical models.

Keywords

Asymmetric information; Bose–Einstein distribution; Contagion; Motion pictures, economics of; Optimal contracts; Pareto distribution; Power laws; Superstars, economics of

JEL Classifications

L82; L20; Z10; D40

The business of motion pictures is a fascinating laboratory for applied researchers in the social sciences. The glamorous subject matter makes the industry inherently interesting, but more important for empirical research is the availability of project-level data on investment and financial returns. Most studies of investment decisions are conducted at the industry or firm level, so that the researcher observes the return only on a portfolio of projects. In the movie business, the unit of observation is the individual project, and data are collected and reported in fine detail by many industry sources.

Early research on the movie business applied microeconomic theory to the industry and made little use of its detailed data and rich institutions. This early literature is important in providing the historical context in which many of the movie industry's business practices emerged. Kindem (1982) has collected in his volume many papers that provide organizational and institutional analyses of the motion-picture industry from its origins through the modern era. More recent papers in this line of applied research provide revisionist analyses of the industry's history and development (Chisholm 1993, 1997; De Vany and Eckert 1991; De Vany and McMillan 2004; Sedgwick 2000).

The market for motion pictures is difficult to understand quantitatively, though the intuition is transparent. Film-goers discover the films they like by consuming them, and through the exchange of information the demand for motion pictures evolves over time. Supply adjusts as the available screens respond to demand through flexible state-contingent exhibition contracts. The present article provides an overview of the economics of motion pictures. The focus is on the how the demand process affects the distribution of outcomes, how the distribution of outcomes can be quantified, and how this relates to the industry's organization and business practices.

Movie-Goer Choices and Outcome Uncertainty

Understanding demand is essential if one is to make sense of the movie industry's contracts and business practices. Early viewers of a movie affect the choices of potential viewers – behaviour that goes under the names of herding, contagion, network effects, bandwagons, path-dependence, momentum, and information cascades. The particular models differ in their details, but they are dynamic in that demand depends on revealed demand, or more generally on how group behaviour arises from the interaction of individual decision-makers (Epstein and Axtell 1996). Initial advantages in movie attendance can lead to extreme differences in outcomes when demand has recursive feedback. De Vany and Walls (1996) showed that box-office revenues have a contagion-like property where the week-to-week change in demand is stochastically dependent on previous demand. A big opening of a bad movie can kill it but a big opening of a good movie can lead to an avalanche of attendance and large revenues. Let's examine the demand for movies more closely to see the origins of extreme success and failure.

Assume for simplicity initially that there was only one movie that could be viewed by one consumer at any one time. Consumers choose in random sequence whether or not to go to the movie. If we further assume that the consumers have a common prior belief about the film's quality, then there is a common probability p that a randomly chosen person will choose to see the film. If we let X be the number attending the film, then X is a binomial random variable; it follows

that when consumers share a common prior the film's revenues would follow a binomial distribution. When quality is unknown and priors over quality differ among viewers, p is a random variable. By conditioning on p and integrating over the binomial distribution, we see that each of the $n + 1$ possible outcomes is equally likely; adding uncertainty to the priors transforms the distribution of revenue from the binomial to the uniform distribution.

Now consider information sharing, as has been modelled by Jovanovic (1987), where potential consumers can use information revealed during a film's run to refine their prior on its quality; this sort of information includes the opinions of other viewers, such as expert reviewers, advertising, and information from box office reports and queuing at cinemas. De Vany and Walls (1996) let the distribution of customers over screens be multinomial uniform, so the movie search problem – a search for quality with an unknown distribution – is similar to the search for price with an unknown distribution. Viewers who do not know the distribution begin with a uniform prior and adapt from there. The result of this process is the Bose–Einstein distribution which has the property that all of the possible outcome vectors are equally likely! This means the vector in which the attendance at every theatre is equal to zero is as likely as one in which all n trials go to only one theatre and every other vector is equally likely (Feller 1957). The Bose–Einstein distribution has uniform mass over a space of $s + 1$-vectors; the s-vectors correspond to the revenues of the s theatres and one bin collects those who go to no film.

What is important about the evolution of choice probabilities under the Bose–Einstein choice logic is the way past successes are leveraged into future successes: as soon as individual differences emerge among the films, they are compounded by information feedback into very large differences over the course of a film's theatrical lifetime. A broad opening at many theatres can produce large and rapidly growing audiences, but it also can lead to early failure if the large crowd relays negative information. Movie customers sequentially select movies, and the probability that a given customer selects a particular movie is proportional to the fraction of previous customers who selected that movie. This result obtains because the probabilities are not known and sampling reveals information that causes previous selections to attract new ones.

Quantifying the Distribution of Movie Outcomes

Box-office revenue is asymptotically power law or Pareto distributed (De Vany and Walls 1999, 2002). One of the attractions of the power law distributions in explaining the movie business is that they allow for the heavy tails and skewness that are characteristic of box-office outcomes. Power laws emerge in many other systems with feedback of the type discussed above (Brock 1999).

The Stable Paretian Model

Mandelbrot (1963) proposed the stable Paretian distribution as a general model for natural and social systems; it is applied in economics, finance, biology, geology, physiology, and other sciences (McCulloch 1996; Uchaikin and Zolotarev 1999; Mantegna and Stanley 1995; Levy and Soloman 1997). The stable distribution is the limiting distribution of all stable processes so that it contains the other well-known stable distributions (Cauchy, Lévy, Gaussian) as special cases. Motion picture profit is well fit by a stable distribution with infinite variance and positive skew (Walls 2000; De Vany and Walls 2004). The stable distribution's ability to capture the empirical regularities found in motion picture data and the distribution's statistical foundation on the most general form of central limit theorem make it a natural model of motion picture outcomes. The theoretical reason for thinking that a stable distribution might apply to motion pictures is that Mandelbrot (1963) showed that a dynamic process that is stable under choice, mixture, and aggregation converges in distribution to the stable distribution. If motion picture revenues and costs are discrete time processes with stable increments, then profit will converge to a stable distribution.

Conditional Stable Distribution

In empirical studies it is possible to model the stable Paretian distribution of movie outcomes conditional on a vector of explanatory variables with the use of McCulloch's (1998) stable regression model in which the index of stability α and the regression coefficients are estimated jointly. The stable regression model has the $j¼1$ familiar form of a linear regression $y_i = \beta_0 + \sum_{j=1}^{k} \beta_{ij} x_i + \varepsilon_i$ where the β's are the coefficients to be estimated and the x's are the regressors, but the random disturbance term is assumed to follow a stable distribution with median zero. Estimation of the stable regression model results in an estimate of the regression coefficient β's as well as an estimate of the characteristic exponent α. The regression coefficients in this model represent what is known about the correlates of film success while at the same time permitting the variance of film success at the box office to be infinite. Estimates of this model show that the distribution of returns conditional on a movie's attributes has infinite variance and that returns to production budgets are substantially larger and returns to stars substantially lower than one would estimate using an improperly specified least-squares model (Walls 2005b).

Stretched Exponentials

Concavity in log-log plots of size against rank, also known as a parabolic power law, are interpreted as evidence of increasing returns to information in the demand for motion pictures (De Vany and Walls 1996; Walls 1997; Hand 2001). Frisch and Sornette (1997) propose a multiplicative stochastic process that can explain the deviation of the data relative to a power law distribution, and Sornette (1998) provides rigorous technical details on multiplicative processes leading to power laws and stretched exponentials. Walls (2005a) finds that the stretched exponential distribution fits motion-picture revenue data remarkably well. The stretched exponential distribution does not truncate the upper tail in its estimates of the probability of a movie earning a larger amount than previous movies. The distribution also accounts for the deviation from the strict Pareto power law in a way that does not place artificial restrictions on the possibility that a movie can earn far more than our experience suggests.

Understanding the Movie Business

We now discuss how the behavioural and statistical models help us to understand the way the motion-picture industry operates and how contracts and business practices adjust the supply of theatrical engagements to capture the increasing returns inherent in the demand process.

The Opening

Stars, large production budgets and national advertising campaigns can place a film on many exhibitor screens when it opens. This can generate high initial revenues and, if viewers like the film and spread the word, it will earn high revenues in the following weeks. But a wide release is vulnerable to negative feedback – if viewers do not like the film, the large opening audience transmits a large flow of negative information, and revenue may decline at a rapid rate. A wide release lowers the gross revenue per theatre, and this may cause exhibitors to drop the film sooner than they would otherwise. The willingness of exhibitors and downstream sources of revenue like cable television, videocassette distributors, pay per view and network television as well as foreign distributors to pay advance guarantees for motion pictures before their theatrical run is a major inducement for distributors to produce big budget films and promote them heavily. The theatrical market can be less important than other sources of revenues (Rusco and Walls 2004).

Decentralization

Each film's run through the market is sequential in order to exploit information dynamics. The run is self-organized because it decentralizes the decision to extend the run to each theatre and uses only local information to extend or close the run at each location. The initial release is modified over time through this process, and new engagements can be added subject to prior contractual obligations. These contractual features interact to adaptively capture revenue and generate strongly increasing

returns from highly successful films. When demand has positive feedback, supply responds flexibly to allow some films to become blockbusters.

Admission pricing

Fixed admission prices (across films but within a given customer class or time of day) are a common industry practice. As a result, demand is accommodated by lengthening a film's run. A relatively stationary admission price combined with a count of admissions gives a reliable signal of demand, and this signal is transmitted throughout the industry by real time reporting of box office revenues. This reporting is required in the exhibition contract and encouraged by other means as well. If the admission price were increased to ration excess demand, the number of people who would see the film in the opening weeks would fall and this would reduce the flow of information from this source to potential viewers. This lower rate of information transfer would lead to a shorter run and a lower total level of demand. The ability to extend the run makes an almost perfectly elastic supply response possible, so there is no need for price to rise to ration excess demand. Fixed admission prices lead to a pure quantity signal and an adaptive supply response to accommodate demand discovery.

Contracting

Optimal contract theory does not fit the environment of motion pictures where expected values are dominated by the rare and unpredictable events that are so large. The incentive clauses of optimal contract theory are designed to alter the probabilities of favourable outcomes and raise expected values, but the asymmetric information often emphasized by optimal contract theory is not a factor because both principal and agent are in a state of symmetric ignorance about the prospects of a movie owing to the 'nobody knows' property (Caves 2000).

A difficult problem to solve contractually is how to keep a film on screens long enough for it to build an audience. If an exhibitor takes such a film, it is with the risk that it may build so slowly during his or her run that only exhibitors who show it later will benefit from information feedback. Because the Paramount decrees bar long-term, exclusive showings, it is difficult to guarantee that the exhibitor who takes the risk of introducing the film will benefit if the film later becomes a success (De Vany and Eckert 1991). When the audience grows recursively, the Paramount contracting restrictions may prevent risk-taking exhibitors from capturing the demand externality which they create.

Extreme events drive the business, so contracts condition pay on rare events with compensation related to the outcome of the movie. Many Hollywood superstar movie contracts contain some form of profit participation (Weinstein 1998). Many contracts are contingent on theatrical box office revenues, which are readily monitored. In this case, the share of gross revenue paid often is nonlinear, with the share rising at higher outcomes, to reflect the nonlinear dependence of profit on revenue. In a complex contract, there may be several breakpoints where the star's percentage share increases, this nonlinearity reflecting the nonlinearity of profit in revenue.

Film Rentals

The exhibition contracts are rich in contingencies that make them highly adaptive: they rely on locally generated information; they set the rental fee in a precise and nonlinear way in response to demand; they share risk between exhibitors and distributors; and they create incentives for exhibitors to show films by granting a measure of exclusivity. The rental price adapts to the state of demand and the rental schedule is nonlinear. Events in the tail are the high-revenue weeks during a movie's theatrical run, and these weeks can occur at any time during the run. During these high-revenue weeks, the rental clause allows the exhibitor to retain his or her (negotiated) cost per week of operation plus ten per cent while allocating the remaining 90 per cent to the distributor (De Vany and Eckert 1991).

Star Power

Movies with superstars have a different distribution of profit from other movies (De Vany and Walls 2004). The profit distribution for superstar movies is an asymmetric stable distribution with

infinite variance. Stars place much more mass in the upper tail of the profit distribution. The probability of extreme catastrophes – losses in excess of $95 million, say – is higher for movies *without* stars than for movies with stars. This is not at all obvious and may not be observed in a given sample. Putting a star in a movie places more mass on the upper tail and less on the lower tail.

Expected profit is positive for star movies and negative for non-star movies. These values are consistent with the fact that probability is skewed to the positive tail in superstar movies and to the negative tail for others. Superstar movies are more profitable and less risky than other movies.

Success Breeds Success

An interesting property of the stable Paretian distribution discussed above is that conditional expectation does not converge. The tails of stable distributions are Paretian and the conditional probability that $x \geq x_0$ is $P[x > x_0] = (x_0 = x)^{\alpha}$. The conditional mean, given that $x > x_0$ equals $\bar{x}_{x0} = x_0\alpha/(\alpha - 1)$. Since α is a constant, the conditional expected value of profit depends linearly on x_0. Conditional on having earned a profit, the expected profit continues to rise with current profit, and this does not end as the movie earns more profit. This is not paradoxical because movies that make it into the upper tail of the profit distribution have been selected from among their competitors. The heavy tails of the stable distribution imply that probability does not decline rapidly enough for the conditional expectation to converge. For the Gaussian or log Gaussian distributions, the conditional expectation converges to a constant as the conditioning event increases. The linear conditional expectation of the Paretian distribution means that blockbuster movies that have already attained high profit have an expectation of even higher profit, and this prospect does not diminish as profit grows. This captures the idea of demand momentum.

Conclusion

When movie audiences see a movie they like, they make a discovery and they tell their friends about it. This and other information is transmitted to other consumers, and demand develops dynamically as the audience sequentially discovers which movies it likes. Supply adapts to revealed demand through flexible exhibition contracts and other business practices that permit the increasing returns in film demand to be realized.

See Also

- ▶ Information Cascades
- ▶ Pareto Distribution
- ▶ Path Dependence
- ▶ Power Laws
- ▶ Superstars, Economics of

Bibliography

Brock, W. 1999. Scaling in economics: A reader's guide. *Industrial and Corporate Change* 8: 403–446.

Caves, R. 2000. *Creative industries: Contracts between art and commerce*. Cambridge, MA: Harvard University Press.

Chisholm, D. 1993. Asset specificity and long-term contracts: The case of the motion-pictures industry. *Eastern Economic Journal* 19: 143–155.

Chisholm, D. 1997. Profit-sharing versus fixed-payment contracts: Evidence from the motion pictures industry. *Journal of Law, Economics, and Organization* 13: 169–201.

De Vany, A., and R. Eckert. 1991. Motion picture antitrust: The Paramount cases revisited. *Research in Law and Economics* 14: 51–112.

De Vany, A., and H. McMillan. 2004. Was the antitrust action that broke up the movie studios good for the movies? Evidence from the stock market. *American Law and Economics Review* 6: 135–153.

De Vany, A., and W.D. Walls. 1996. Bose–Einstein dynamics and adaptive contracting in the motion picture industry. *Economic Journal* 106: 1493–1514.

De Vany, A., and W.D. Walls. 1999. Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of Cultural Economics* 23: 285–318.

De Vany, A., and W.D. Walls. 2002. Does Hollywood make too many R-rated movies?: Risk, stochastic dominance, and the illusion of expectation. *Journal of Business* 75: 425–451.

De Vany, A., and W.D. Walls. 2004. Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. *Journal of Economic Dynamics and Control* 28: 1035–1057.

Epstein, J., and R. Axtell. 1996. *Growing artificial societies: Social science from the bottom up*. Cambridge, MA: Brookings Institution and MIT Press.

Feller, W. 1957. *An introduction to probability theory and its applications*. New York: Wiley.
Frisch, U., and D. Sornette. 1997. Extreme deviations and applications. *Journal de Physique* 1: 1155–1171.
Hand, C. 2001. Increasing returns to information: Further evidence from the UK film market. *Applied Economics Letters* 8: 419–421.
Jovanovic, B. 1987. Micro shocks and aggregate risk. *Quarterly Journal of Economics* 17: 395–409.
Kindem, G. (ed.). 1982. *The American movie industry: The business of motion pictures*. Carbondale: Southern Illinois University Press.
Levy, M., and S. Soloman. 1997. New evidence for the power law distribution of wealth. *Physica A* 242: 90–94.
Mandelbrot, B. 1963. New methods in statistical economics. *Journal of Political Economy* 71: 421–440.
Mantegna, R., and H. Stanley. 1995. Scaling in financial markets. *Nature* 376: 46–49.
McCulloch, J. 1996. Financial applications of stable distributions. In *Statistical methods in finance, vol. 14 of handbook of statistics*, ed. G. Maddala and C. Rao. New York: North-Holland.
McCulloch, J. 1998. Numerical approximation of the symmetric stable distribution and density. In *A practical guide to heavy tails: Statistical techniques and applications*, ed. R. Adler, R. Feldman, and M. Taqqu. Berlin: Birkhäuser.
Rusco, F., and W.D. Walls. 2004. Independent film finance, pre-sale agreements, and the distribution of film earnings. In *The economics of art and culture*, Contributions to Economic Analysis, vol. 260, ed. V. Ginsburgh. Amsterdam: Elsevier.
Sedgwick, J. 2000. *Popular filmgoing in 1930s Britain: A choice of pleasures*. Exeter: University of Exeter Press.
Sornette, D. 1998. Multiplicative processes and power laws. *Physical Review E* 57: 4811–4813.
Uchaikin, V., and V. Zolotarev. 1999. *Chance and stability: Stable distributions and their applications*. Utrecht: VSP.
Walls, W.D. 1997. Increasing returns to information: Evidence from the Hong Kong movie market. *Applied Economics Letters* 4: 187–190.
Walls, W.D. 2000. Measuring and managing uncertainty with an application to the Hong Kong movie business. *International Journal of Management* 17: 118–127.
Walls, W.D. 2005a. Demand stochastics, supply adaptation, and the distribution of film earnings. *Applied Economics Letters* 12: 619–623.
Walls, W.D. 2005b. Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns. *Journal of Cultural Economics* 29(3): 177–190.
Weinstein, M. 1998. Profit-sharing contracts in Hollywood: Evolution and analysis. *Journal of Legal Studies* 27: 67–112.

Müller, Adam Heinrich (1779–1829)

Hermann Reich

Born in Berlin, Müller studied in Göttingen and became a private tutor and scholar. In 1811 he had to leave Berlin because of his opposition to the reforms of Hardenberg, and later served the Austrian foreign minister Metternich in various – partly conspiratorial – positions, for which he was created knight of Nitterdorf in 1826. An ardent catholic – he had converted in 1805 – Müller opposed the ideals of the Enlightenment, and rejected liberalism, rationalism, individualism and materialism. He was a bitter enemy of the French Revolution and one of the intellectual voices of the post-Napoleonic restoration.

Müller was the most important political economist of the German Romantic school. A central element in his thinking was the organic unity of society and state. The society and its economy constitute an organic totality which is more than the sum of the economies of its individual members. This totality is represented by the state, which is an end in itself (Müller 1808–9, book 1; 1816, pt 1). Thus he opposed the economic theories of Adam Smith and his successors, particularly their abstract and isolated understanding of the individual and their emphasis on self-interest. He also criticized Smith's merely materialist notion of national wealth and formulated a concept of spiritual capital encompassing cultural values and the state of the sciences (Müller 1808–9, books 4 and 5; 1931, ch. 10). Beside his other works he developed an interesting sociological theory of money (Müller 1816, pt 2).

Müller was anti-capitalist and anti-industrialist and one of the first to raise – though rudimentarily – the social question. He regarded the social organization of the Middle Ages, with its traditional hierarchy and its guilds, as a model for his reactionary utopia. His political impact was very limited: his extreme anti-modernism had to collide with conservative

Realpolitik. Nevertheless, Müller's eminent role in the articulation of German right-wing anti-capitalism was to remain of lasting importance.

Selected Works

1808–9. *Die Elemente der Staatskunst*. Berlin: Haude & Spenersche Verlagsbuchhandlung, 1936.
1816. *Versuche einer neuen Theorie des Geldes mit besonderer Rücksicht auf Grossbritannien*. In ed. H. Lieser. Jena: Gustav Fischer, 1922.
1923. *Schriften zur Staatsphilosophie*. In ed. R. Kohler. Munich: Theatiner.
1931. *Ausgewählte Abhandlungen*. In ed. J. Baxa. Jena: Gustav Fischer.

Bibliography

Baxa, J. 1930. *Adam Müller: Ein Lebensbild aus den Befreiungskriegen und aus der deutschen Restauration*. Jena: Gustav Fischer.

Multicollinearity

Wilfred Corlett

Exact multicollinearity means that there is at least one exact linear relation between the column vectors of the $n \times k$ data matrix of n observations on k variables. More commonly, multicollinearity means that the variables are so intercorrelated in the data that the relations are 'almost exact'. The term was used by Frisch (1934) mainly in the context of attempts to estimate an exact relation between the systematic components of variables whose observed values contained disturbances or errors of measurement but where there might also be other relations between the systematic components which made estimates dangerous or even meaningless. In more recent work the data matrix has usually been the matrix X of regressor values in the linear regression model $Y = X\beta + \varepsilon$ with no measurement errors. Confusion between the two cases led at one time to some misunderstanding in the literature. Other terms used for the same phenomenon are collinearity and ill-conditioned data – although the latter may contain aspects of the scaling of variables which are irrelevant to multicollinearity.

For the linear regression model, exact multicollinearity means that it is impossible to separate out the effects of the individual variables in an exact relation. Some (or all) of the parameters of the model can not be estimated although some linear functions of them can. However, exact multicollinearity is rare except in badly specified models. Multicollinearity that is not exact is liable to lead to high variances and standard errors for least squares estimators of the parameters, although now some linear functions of the parameters can be estimated with much smaller variances. As a result, tests of hypotheses about the parameters may have little power, so that very inaccurate hypotheses may not be rejected, and confidence intervals for the parameters may be very wide. If a test rejects an hypothesis, multicollinearity has not done any serious harm; if it does not reject it, multicollinearity *may* be the cause but there are other reasons for not rejecting. The sample may be too small; the variance of the error ε may be too high; there may be too little variation in the relevant variable – although this is sometimes considered as collinearity with a constant in the model; or the hypothesis may be correct or almost correct.

Various methods have been suggested for assessing the importance of multicollinearity. One was the use of bunch maps (Frisch 1934) which involved calculating regressions in all possible subsets of the variables, with minimization in the direction of each variable in the subset in turn, followed by a diagrammatic presentation. Interpretation was difficult. Several methods use the matrix R of correlations between the X variables. High off-diagonal elements indicate possible harmful effects from relations between two variables but relations involving more variables are difficult to spot in this way. Following

Farrar and Glauber (1967) attention was paid to the determinant of *R,* with value near zero indicating serious multicollinearity. The diagonal elements of the inverse of *R,* the variance inflation factors, show how much multicollinearity multiplies the variance which would be achieved for the least squares estimator of a parameter if the variable associated with it were orthogonal to all the other variables. Possibly the most fruitful method is the use of the eigenvalues of R or of $X'X$ where each column of X has been scaled to have unit length but has not been centred. Belsley et al. (1980) use this scaled X and the singular values of X (which are the square roots of the eigenvalues of $X'X$) to detect the number of apparently harmful relations between the variables, the effects of each on the estimation of each parameter and, hence, in combination with auxiliary regressions, an indication of which variables they contain.

If multicollinearity causes serious problems in some application, there are various possible ways of improving the situation. It may be possible to obtain more data. If the data really satisfy the model, this will improve matters – particularly if the new data do not have the same collinearities as the old. It may be possible to impose exact restrictions on the parameters. These restrictions, usually derived from economic theory, can give considerable improvements to the properties of estimators, but only if they are really justified. Finally, it may be possible to use stochastic information on the parameters or linear functions of them. This may be done by pure Bayesian techniques, or, following Theil and Goldberger (1961), by a form of mixed estimation using stochastic information on the values of the parameters, or linear functions of them, obtained from previous samples or from introspection (cf. Theil 1971).

See Also

- ▶ Bunch Maps
- ▶ Estimation
- ▶ Multivariate Time Series Models
- ▶ Simultaneous Equations Models

Bibliography

Besley, D.A., E. Kuh, and R.E. Welsch. 1980. *Regression diagnostics*. New York: Wiley.

Farrar, D.E., and R.R. Glauber. 1967. Multicollinearity in regression analysis: The problem revisited. *Review of Economics and Statistics* 49: 92–107.

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

Theil, H. 1971. *Principles of econometrics*. New York: Wiley.

Theil, H., and A.S. Goldberger. 1961. On pure and mixed statistical estimation in economics. *International Economic Review* 2: 65–78.

Multilingualism

Victor Ginsburgh and Shlomo Weber

Abstract

Multilingualism or *linguistic diversity* in a heterogeneous society provides extraordinary challenges and room for policies which may have important economic implications in shaping the flows of interregional or international trade, investment and migrations. Given the often uncompromising nature of linguistic conflicts, linguistic policies and, especially, the choice of official languages should take into account the preferences of those groups of individuals whose cultural, societal, historical values and sensibilities could be affected. In evaluating linguistic policies an important role is played by the dynamic nature of language environments driven by individual choices of learning other languages.

Keywords

Communicative benefits; Linguistic disenfranchisement; Linguistic standardization; Multilingualism; Official languages

JEL Classifications

D63; D70; O52; Z13

Multilingualism or linguistic diversity is an important societal phenomenon that can generate gains or losses resulting from the economic interactions between individuals, regions or countries. The effects of multilingualism have recently come to the forefront of public policy debates. Linguistic issues and, in particular, the treatment of minority languages are almost unparalleled in terms of their explosiveness and emotional appeal, much more so than any other question of resource allocation or responsibility sharing within a polity. As noted by Bretton (1976, p. 447), 'language may be the most explosive issue universally and over time. This mainly because language alone, unlike all other concerns associated with nationalism and ethnocentrism is so closely tied to the individual self. Fear of being deprived of communicating skills seems to raise political passion to fever pitch.'

Language policies in multilingual societies are beset by the trade-off between standardization and disenfranchisement. Linguistic *standardization* comprises any set of policies that promote the dominant use of a unique or several languages while limiting the usage of languages spoken by other population groups. Indeed, linguistic standardization may deliver important benefits in terms of greater ease of communication, reducing costs of translation, increased trade, improved economic performance and administrative efficiency. However, excessive standardization may exacerbate the alienation of large minorities and widen the existing chasm between linguistic communities (Laponce 2003). A restriction of basic linguistic rights may create *disenfranchisement* of groups of individuals and cause citizens to lose their ability to communicate in the language of their choice. Standardization, which is often represented by a selection of official languages and allocation of linguistic rights, may alienate those groups of individuals whose cultural, societal and historical values and sensibilities are not represented by the official languages (Laitin 1989). As Pool (1991) points out, non-official languages may suffer from their 'minority status' and limit employment and advancement possibilities of their native speakers.

Since in many cases it is not feasible to include all the languages in the set of official ones, a multilingual society must design some language standardization policies (for example, the 'three-language formula' in India; Baldridge 1996) and the implementation of certain standardization measures (De Swaan 2001; Grin 2004). However, the explosive and uncompromising nature of linguistic conflicts, the reluctance of linguistic majorities to concede rights to minorities, makes the choice of official languages a challenging and daunting task. Thus, the choice of the set of official languages has to take into account the sensitivity of a society towards possible disenfranchisement of large groups of its citizens (Ginsburgh et al. 2005) and has to rely on a delicate resolution of the interplay between administrative and cost efficiency, on the one hand, and the rights and desires of various linguistic groups, on the other (Van Parijs 2005).

To illustrate the individual and aggregate cost and benefits of standardization and disenfranchisement, we consider a society M and the set of languages L spoken in this society. We assume that every citizen i is endowed with a unique native language $n(i) \in L$ and a set of languages $L(i) \subset L$ that, to simplify, she commands with identical ease. A linguistic profile of each individual i is the pair $(n(i), L(i))$, and society's linguistic profile is given by $P = (n(i), L(i))_{i \in M}$. A linguistic policy is represented by a set of official languages $K \subset L$ that is chosen for administrative, educational, and official communication functions in the society (Pool 1991, 1996, and the extensive list of references therein; Ginsburgh et al. 2005.) The choice of the set K represents a linguistic *standardization policy*. If the set of official languages K is non-empty and smaller than L, those members of the society whose native language is not included in K will be *disenfranchised* and some of their linguistic rights will be denied.

In order to evaluate the costs of disenfranchisement, we assume that every citizen i has utility function u_i defined over all subsets of L. We will denote $u_i(K)$ for $i \text{ A } \in M$ and $K \subset L$, where citizens with the same linguistic profiles have identical utility functions. It is important to stress that the functions u_i are defined over the set of languages as a whole, rather than being dissected into preferences over single languages. Though citizens may have preferences over single

languages, their evaluation of the set of official languages could be crucially affected by inclusion or exclusion of their native language. The aggregate utility (welfare) function for the entire society is given by $W(u, P, K)$, where u is the vector of u_i's.

Our description indicates the special role played by the native languages of citizens in M, which can be viewed as the union of linguistic clusters M_l, where, for each $l \in L$, M_l consists of citizens whose native language is l. Assuming additivity of the aggregate utility, we have $W(u,P,K) = \sum_{l \in L} \sum_{l \in M_l} u_i(K)$. As a simple example, consider the *dichotomous* function based on the citizens' native languages (Ginsburgh and Weber 2005), for which the value of $u_i(K)$ is 1 if i 's native language, $n(i)$, is included in K, and zero if it is not. The latter group contains individuals who are *disenfranchised* by the imposed standardized measures. The value taken by the function W is the number of citizens whose native language belongs to the set K, $W^1(u,P,K) = \sum_{\{i \in N \mid n(i) \in K\}} 1$. One generalization of the dichotomous approach is to take into account the entire language profile of every citizen rather than her native language only. Then, the value of her utility function is 1 if at least one of the languages spoken by her is included in K and zero otherwise. Here, the notion of disenfranchisement is limited to those who speak no official language: $W^2(u,P,K) = \sum_{\{i \in N \mid L(i) \cap K \neq \varnothing\}} 1$.

In evaluating citizens' preferences over subsets of languages one may take into account the similarity or the proximity between languages (see, for example, Dyen et al. 1992, for a matrix of distances between 95 Indo-European languages). Let $\delta(l,l')$ be the linguistic distance between two languages l and l'. Denote the linguistic distance between any two subsets T, T' of L as the minimal distance between a language from T and a language from T': $\delta(T,T') = \min_{l \in T, l' \in T'} \delta(l,l')$. Then, the 'linguistic welfare' of the society is function of the distances between citizens' native languages and the set of official languages K : $W^3(u, P, K) = w(\delta(n(1), K), \delta(n(2), K), \ldots, \delta(n(M), K))$, where $w : \mathcal{R}_+^M \to \mathcal{R}$ is decreasing in each of its M arguments. Again, a modified utility function could be defined over the distances between the sets $L(i)$ and K instead: $W^4(u, P, K) = w(\delta(L(1), K), \delta(L(2), K), \ldots, \delta(L(M), K))$.

Note that enlarging the set of official language is welfare improving in all four specifications above. Thus, if the only goal of the society is to maximize aggregate utility, it should set $K = L$. However, there are also other considerations to take into account. Difficulties of communication, costs incurred by translation and interpretation, possible errors causing delays and sometimes paralysing multilateral discussions and negotiations impose a non-negligible burden on societies with a large number of official languages (in 2007, the European Union had to manage 23 official languages at a cost over $1.5 billion). Denote then by $C(K)$ the cost of maintaining the set K of official languages. Obviously, C is increasing, but its specific form depends on the intensity of the linguistic regime. There could be various requirements, including a 'full' regime that every official document needs to exist in all official languages.

There is thus a trade-off between language standardization (and disenfranchisement of some citizens) and the translation, interpretation and communication costs generated by every additional official language. Formally, the society's objective is to find a set of languages K that maximizes the difference between aggregate utility and costs: $\max_{K \subset L} W(u,P,K) - C(K)$. A solution to this problem is discussed by Grin (2004, p. 201), who argues that there must be an optimum, since 'it is reasonable to assume that the benefits of diversity increase at a decreasing rate, while its costs increase at an increasing rate', and is addressed in Ginsburgh et al. (2005).

Language profiles considered so far are assumed given. In fact, they can be remarkably dynamic and change over time as individuals may decide to learn other languages. The reasons that induce citizens to do so can be analysed by examining the benefits and the costs that such learning generates. Benefits are often linked with the increased earning potential, especially in the case of immigrants who acquire the native language of the country in which they live (see, for example,

MacManus et al. 1978; Grenier 1985; Lang 1986; Chiswick 1998; and references in Grin and Vaillancourt 1997). We consider the Selten and Pool (1991, p. 66) 'communicative benefits' approach that frees itself from the restriction that 'earnings [are] a mechanism and firms a milieu of the incentive to learn languages'. For every language l consider the set M_l of its native speakers, whose number is denoted by m_l.

Assume for simplicity that $L = j$; k and that all citizens speak only their native language, so that the linguistic profile $L(i)$ consists of $n(i)$ for every i A M. Citizens may learn the other language. Denote by $m_{j,k}$,($m_{j,k}$) the number of citizens in $M_j(M_k)$ who do so. A citizen $i \in M_j$ who learns language k incurs a cost $C(\delta(j,k))$, where C is an increasing function of linguistic distance. Let $u_j(m_j,)$ be the utility of $i \in M_j$, where the second argument indicates the number of individuals i can communicate with. We assume that the utility functions are increasing and, moreover, identical for all individuals with the same native language. If i learns k, it costs her $C_{j,k}$, but she will be able to communicate with all citizens in M_k. Her gross benefit will be given by $u_j(m_j, m_k)$. If i does not learn k, she will be able to communicate with those in M_k who learn language j, and her gross (and net) benefit will be $u_j(m_j, m_k)$. This formulation leads to the following equilibrium condition that makes individuals in

M_k indifferent between learning the other language and deciding not to do so: $u_j(m_j; m_k) - C_{j;k} = u_j(m_j; m_{k;j})$. This equation allows us to determine the number of citizens in group M_k who learn j, and in a similar manner the number of those in group M_j who learn k (see Selten and Pool 1991; Church and King 1993; Shy 2001; Gabszewicz et al. 2005; Ginsburgh et al. 2007). By imposing some additional conditions, such as continuity, concavity and supermodularity of the utility functions one can derive some comparative statics results. In particular, one can show that the number of learners of the foreign language j in country k is positively correlated with the number of j-speakers in other countries and negatively correlated with the population size of their own country k (Lazear 1999; Ginsburgh et al. 2007). These results also show that public policies may be useful in stimulating learning (for a cost–benefit analysis of linguistic policies in Quebec, see, for example, Breton and Mieskowski 1975, Vaillancourt 1987; see also Fidrmuc and Ginsburgh 2007, for policy suggestions in the EU).

In short, the questions raised by multilingualism offer serious challenges and the main reason is that linguistic policies are concerned not only with difficult trade-offs and resource allocation issues, but enter also the area of public policies that touch so closely personal values, beliefs and traditions.

See Also

- ▶ Culture and Economics
- ▶ Social Welfare Function

Acknowledgment We should like to thank Yuval Weber for his help in preparing this manuscript.

Bibliography

Baldridge, J. 1996. Reconciling linguistic diversity: The history and future of linguistic policies in India. Discussion paper, University of Pennsylvania.

Breton, A., and P. Mieskowski. 1975. The returns to investment in language: The economics of bilingualism. Working paper no. 7512, Toronto Institute for the Quantitative Analysis of Social and Economic Policy, University of Toronto.

Bretton, H. 1976. Political science, language, and politics. In *Language and politics*, ed. W.M. O'Barr and J.F. O'Barr. The Hague: Mouton.

Chiswick, B. 1998. Hebrew language usage: Determinants and effects on earnings among immigrants in Israel. *Journal of Population Economics* 15: 253–271.

Church, J., and I. King. 1993. Bilingualism and network externalities. *Canadian Journal of Economics* 26: 337–345.

De Swaan, A. 2001. *Words of the world*. Cambridge: Polity Press.

Dyen, I., J.B. Kruskal, and P. Black. 1992. An Indo-European classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5): 1–132.

Fidrmuc, J., and V. Ginsburgh. 2007. Languages in the European Union: The quest for equality and its cost. *European Economic Review* 51: 1351–1369.

Gabszewicz, J., V. Ginsburgh, and S. Weber. 2005. Bilingualism and communicative benefits. Discussion paper, CORE, Catholic University of Louvain.
Ginsburgh, V., and S. Weber. 2005. Language disenfranchisement in the European Union. *Journal of Common Market Studies* 43: 273–286.
Ginsburgh, V., I. Ortuño-Ortín, and S. Weber. 2005. Language disenfranchisement in linguistically diverse societies: The case of European Union. *Journal of the European Economic Association* 3: 946–965.
Ginsburgh, V., I. Ortuño-Ortín, and S. Weber. 2007. Learning foreign languages: Theoretical and empirical implications of the Selten and Pool model. *Journal of Economic Behavior and Organization* 64: 337–347.
Grenier, G. 1985. Bilinguisme, transferts linguistiques et revenus du travail au Québec, quelques éléments d'interaction. In *Economie et Langue*, ed. F. Vaillancourt. Québec: Editeur officiel.
Grin, F. 2004. On the costs of cultural diversity. In *Cultural diversity versus economic solidarity*, ed. F. Van Parijs. Brussels: De Boeck Université.
Grin, F., and F. Vaillancourt. 1997. The economics of multilingualism: Overview of the literature and analytical framework. In *Multilingualism and multilingual communities*, ed. W. Grabbe. Cambridge, MA: Cambridge University Press.
Laitin, D. 1989. Language policy and political strategy in India. *Policy Sciences* 21: 415–436.
Lang, K. 1986. A language theory of discrimination. *Quarterly Journal of Economics* 100: 363–381.
Laponce, J.A. 2003. Minority languages and globalization. *Nationalism and Ethnic Policies* 10: 15–24.
Lazear, E. 1999. Culture and language. *Journal of Political Economy* 107: 95–126.
MacManus, W., W. Gould, and F. Welsch. 1978. Earnings of Hispanic men: The role of English language proficiency. *Journal of Labor Economics* 1: 101–130.
Pool, J. 1991. The official language problem. *American Political Science Review* 85: 495–514.
Pool, J. 1996. Optimal language regimes for the European Union. *International Journal of Sociology of Language* 121: 159–179.
Selten, R., and J. Pool. 1991. The distribution of foreign language skills as a game equilibrium. In *Game equilibrium models*, ed. R. Selten, Vol. 4. Berlin: Springer-Verlag.
Shy, O. 2001. *The economics of network industries*. Cambridge: Cambridge University Press.
Vaillancourt, F. 1987. The benefits and costs of language policies in Quebec, 1974–1984: Some partial estimates. In *The economics of language use*, ed. H. Tonkin and K. Johnson-Weiner. New York: Center for Research and Documentation on World Language Problems.
Van Parijs, P. 2005. Europe's three language problems. In *The challenge of multilingualism in law and politics*, ed. D. Castiglione and C. Longman. Oxford: Hart Publishing.

Multinational Corporations

Edith Penrose

Abstract

After World War II economists began to notice that direct private foreign investment seemed to be increasingly associated with the expansion of very large firms, mostly, but not entirely, form the United States and that this phenomenon was attracting considerable political criticism. Some economists, early called 'institutionalists', had long been concerned with the study of the firm as an economic organization but the main stream of economic theorists had paid scant attention to it, concentrated as they were on the theory of prices and the allocation of resources.

After World War II economists began to notice that direct private foreign investment seemed to be increasingly associated with the expansion of very large firms, mostly, but not entirely, form the United States and that this phenomenon was attracting considerable political criticism. Some economists, early called 'institutionalists', had long been concerned with the study of the firm as an economic organization but the main stream of economic theorists had paid scant attention to it, concentrated as they were on the theory of prices and the allocation of resources.

The development of the theory of monopolistic competition had introduced a new and important line of thought in the 1930s but although it continued to flourish on its own, there seemed no satisfactory way of incorporating it into 'general' economic theory. An early article by R.H. Coase (1937) explaining the limits to what the market could efficiently do and why firms existed at all was widely cited and praised as a 'seminal' contribution but nothing much followed from it. Such early and broadly based economists as

Alfred Marshall, J.B. Clark, Frank Knight and D.H. Robertson had paid some attention to the fact that production was organized by firms, but it was more or less generally agreed that its location and composition, relative prices and the allocation of goods and services were determined by market forces. Economists continued to treat direct private foreign investment within the traditional framework of the pricing system and thus simply as capital flows determined by international differences in the rate of return on capital. The role of the firm did not come within the purview of economic theory except as an aspect of imperfect (monopolistic) competition. In any case, the size and growth of firms continued to be regarded as confined by rising cost and/or falling demand curves.

Coase had departed from this orthodoxy by demonstrating that the administrative organization of production within the firm often had what are now called 'transactional advantages' over the market. Penrose (1959) in her study of the growth of the firm enquired into the process of growth and its limits (if any). Using very different terminology, she demonstrated that there were often transactional advantages for the firm when it made use of its owned internal resources, some of which could not be bought in the market but by their very nature were produced only within the firm in the process of growth. She also concluded that although there was no limit to the size of a firm so long as it could be administered as a coherent entity, there was a limit to its rate of growth, but that as a firm grew there was no evidence that its administrative capacity could not grow accordingly.

Hymer, in his MIT thesis (1960) was the first to challenge directly the received theory of direct private foreign investment by convincingly demonstrating its inability to explain the type or geographical distribution of capital flows. Moreover, the type of companies involved with direct foreign investment seemed to be motivated primarily by competitive conditions in particular markets rather than by differences in rates of interest even if properly adjusted for differences in risk. With Hymer's work the floodgates of theoretical and empirical studies of the multinational corporation (MNC) were opened and economists began to enquire into what difference it made when direct investment took place in a variety of fields and areas under the coordinating umbrella of a large administrative organization.

Predictably, economists began to adopt new terminology in an attempt to indicate that this type of investment was not to be regarded as simply an international movement of capital but as the movement of a bundle of resources; was undertaken by firms that were not international in ownership but were usually national firms operating in a number of countries through separately incorporated enterprises connected with and responsible to a central headquarters; and, above all, was administratively organized on such a scale as to displace the 'market' over wide and varied types of activity. By the early 1980s many such firms were increasingly referred to as 'global' corporations forming and implementing competitive strategies on a 'global' scale. In effect, it was recognized that the emergence and growth of MNCs can be regarded essentially as the growth of firms (defined as administrative organizations) through investment abroad.

There was nothing new about this: as early as 1815 a European manufacturer of textile machinery (Cockerill of Belgium) put up a plant in Prussia, 37 years before the first American direct investment (Colt's firearm factory in Britain in 1852). European direct investment in American agriculture was important in the 19th century and American investment abroad grew rapidly. By 1897 Europeans were already complaining of an American invasion; by 1914 the book value of US private foreign investment accounted for 7 per cent of US GDP, the same percentage as in 1966 (Wilkins 1970, p. 202).

American firms introduced in this period a number of new (or superior), often laboursaving, products (e.g. sewing machines, farm machinery, cash registers, elevators, firearms, steam pumps, the telegraph and telephone, locomotives) most of which seemed to have been appropriate responses

to the economic conditions in this newly industrializing country. Furthermore, American firms had been expanding nationally across the continent and this provided experience in dealing with many geographically extensive marketing, managerial and risk-taking problems very similar to those encountered when expanding abroad. Accompanying all this, and essential to it, was the rapid technological development in communications and transportation and in the art and technology of managerial administration and coordination.

Nevertheless, until virtually destroyed by the depression of the 1930s, portfolio investment, largely from Britain, dominated international capital flows. But in the interwar period direct investment increased rapidly as a proportion of total investment and the MNC began to move to the centre of the stage. The common strand in the line of thought from Coase to Hymer lay in the notion of the firm as an internal market for transactions that would not have taken place in external markets and as a more profitable method of conducting some transactions that might otherwise have been at arm's length. If such an internal market has transaction advantages for the firm it implies either pure 'market failure' or imperfect markets for other reasons. Such considerations do not, however, distinguish the MNC from domestic firms and the question arises whether a theoretical distinction is required once it is recognized that there had been no accepted theory of the growth of firms as organizations even within national boundaries, let alone on an international scale.

The same issue had early been raised with respect to international and inter-regional trade. Trade was looked on as a mechanism that tended to equalize factor prices internationally, just as the international firm was looked on as engaging in international capital arbitrage that tended to equalize interest rates. Over fifty years ago Bertil Ohlin (1933) succinctly analysed the nature of the differences between inter-regional and international trade, stressing differences in the quality of productive factors in different countries, the possibility of using entirely different technical processes, and the economies of large-scale production. By and large these are the same types of consideration that influence the investment of MNCs.

The enduring strength of any large firm lies in the quality of its 'owned' productive resources. Of these, perhaps the most important are the knowledge and experience of its personnel, for which the market is inherently as well as institutionally imperfect, its organizational capacity to formulate and implement strategies to utilize such resources, and the marketing advantages conferred by a long-established reputation. Technological expertise in production of goods and/or services, together with the monopoly power thus conferred, was in the first instance seen as the crucial element, but later the management skills required to develop new forms of administrative structures in order to maintain the efficiency of administration as firms grew larger and controlled increasingly diversified activities was seen to be equally important.

Both the managerial structure and the managerial function have undergone fundamental changes that have profoundly affected the nature of the organization itself. Chandler's history (1962) of large American industrial enterprises superbly showed the relationship between growth, strategy and structure. On a more general level the work of Williamson (1975) had especial influence. He analysed the ways in which firms can expand their ability to manage growth efficiently with minimal interference with on-going operations, and he outlined the types of structural change that permit it to avoid rising managerial inefficiency and to develop and implement coherent strategies in a world where competition in technological innovation and in the ability to influence consumer demand are often more prevalent than price competition. The 'organizational theorists' of the firm developed the advantages to be obtained by a firm when production, marketing, research and development, and financial management could be linked and formed into a coherent network of activities guided by a coherent administrative organization with the objective of capitalizing the rents generated at each level.

Most of the economic research on the MNC, therefore, tended to focus on a variety of specific

empirical questions: why do firms prefer to produce abroad rather than export or grant licences to others for the use of their technology? Why do they locate where they do? Why do they have advantages over local firms? Under what circumstances do they integrate horizontally or vertically? In what industries are they most important and why? What countries tend to produce most MNCs, and why? And what kinds of changes in the distribution of source countries occur over time, and why? Empirically, economists were primarily concerned with testing the applicability of the theoretical answers to such questions and in the process, as is normal, they continually produced additional hypotheses.

Nearly all of this work relied heavily on an analysis of the transaction advantages of a 'market' internal to the MNC, some of them of the traditional monopolistic variety and some arising from 'failure' inherent in the market for knowledge and other public goods or as a result of uncertainty. From an economic point of view there are differences between national and international firms but the differences are not such as to require a theoretical distinction between the two types of organization, only a recognition that national boundaries make an empirical difference to their opportunities and costs. Dunning (1981) implicitly recognized this when, eschewing a 'general' theory, he produced what he called an 'eclectic' theory of international production which drew on several branches of economics relating to location, transactional and ownership advantages and the nature of monopolistic competition.

The role of the MNC has also figured prominently in development economics, a branch of economics that has grown rapidly since World War II. The most commonly used measures of economic development are national income per capita, rates of growth, changes in the sectoral distribution of activity, and distribution of income (including employment). All of these may be very much affected by the operations of MNCs in both home and host countries. There seems to be widespread agreement among economists that the important policy issues raised for home countries are very similar to those raised by any outflow of capital or technology and that restrictions on the expansion of their firms abroad may have results similar to those of protectionist trade measures.

There is much more disagreement over the problems raised for host governments in developing countries, not only because they are assumed to be politically less able adequately to prevent monopolistic and certain other deleterious practices, particularly tax evasion through transfer pricing but also because of the effect of the introduction of 'inappropriate' technology on the structure of production and of 'inappropriate' tastes on the composition of demand. Ill-advised policies in both developing and developed countries are often the principal source of the disadvantages created by MNCs for both groups. Caves (1982) presents a good survey of these complex issues and of the sources of disagreement among economists although he ignores the Marxist approaches (but see Warren, 1980).

Partly for the above reasons and partly because they are frequently regarded as instruments of foreign domination, MNCs have often been attacked, restricted, excluded and even expropriated in a large number of developed as well as developing countries. They may come to dominate not only important industries (ranging from mining and oil, to pharmaceutical, electronic and other high technology industries) but at times also the entire economy of small countries dependent on a narrow range of exports. The political as well as economic power they sometimes possess may be, and undoubtedly at times is, used to the disadvantage of host countries. Dislike and fear of foreign domination and of 'monopoly capitalism' generally not infrequently coverage in one many-sided attack on MNCs individually and collectively. Large economic institutions inevitably have a political significance which cannot be ignored in considering their overall impact on any society.

Finally, what can be said about the implications of the 'theories' of the large modern corporation, including the MNC, for the usefulness of the traditional theory of competitive markets? Attempts are frequently made to 'reconcile' or

put into one 'general' theory these two types of approach to the economic world. Such attempts are misguided since the chief function of the latter is to provide a standard against which to judge the effects of the former on prices and on the allocation of resources in the economy. It is equally misguided, however, to interpret the latter as providing a standard of 'welfare' in a changing world characterized by great inequality and by processes akin to Schumpeter's 'creative destruction'. To indiscriminately label deviations from the standard as 'distortions' inimical to welfare in real economies is not just misleading, it is wrong.

See Also

- ▶ Corporate Economy
- ▶ Finance Capital
- ▶ Imperialism
- ▶ Monopoly Capitalism
- ▶ North–South Economic Relations

Bibliography

Bergsten, C.F., T. Horst, and T.H. Moran. 1978. *American multinationals and American interests*. Washington, DC: Brookings Institution.

Buckley, P.J., and M. Casson. 1976. *The future of multinational enterprise*. London: Macmillan.

Caves, R.E. 1982. *Multinational enterprise and economic analysis*, Cambridge surveys of economic literature. Cambridge: Cambridge University Press.

Chandler, A.D. Jr. 1962. *Strategy and structure: chapters in the history of the American industrial enterprise*. Cambridge, MA: MIT Press.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Doz, Y. 1986. *Multinational strategic management*. Oxford: Pergamon Press.

Dunning, J.H., eds. 1971. *The multinational enterprise*. London: George Allen & Unwin.

Dunning, J.H. 1981. *International production and the multinational enterprise*. London: George Allen & Unwin.

Frank, I. 1980. *Foreign enterprise in developing countries*. Baltimore and London: Johns Hopkins Press.

Franko, L.G. 1976. *The European multinationals: A renewed challenge to American and British big business*. Stamford: Greylock.

Garnaut, R., and A. Clunies Ross. 1975. Uncertainty, risk aversion and the taxing of natural resource projects. *Economic Journal* 85: 272–287.

Helleiner, G.K. 1979. Transnational corporations and trade structure: The role of intra-firm trade. In *On the Economics of Intra-industry Trade: Symposium 1978*, ed. H. Giersch. Tübingen: Mohr.

Hymer, S.H. 1960/1976. The international operations of national firms: A study of direct foreign investment. Ph.D. dissertation, Cambridge, MA: MIT Press.

Kindleberger, C.P., eds. 1970. *The international corporation: A symposium*. Cambridge, MA: MIT Press.

Kojima, K. 1975. *Direct foreign investment: A Japanese model of multinational business operations*. New York: Praeger.

Lall, S., and P. Streeten. 1977. *Foreign investment, transnationals and developing countries*. London: Macmillan.

Mikesell, R.F. 1971. *Foreign investment in petroleum and mineral industries: Case studies of investor host-country relations*. Baltimore: Johns Hopkins Press.

Ohlin, B. 1933. *Interregional and international trade*. Cambridge, MA: Harvard University Press.

Penrose, E.T. 1956. Foreign investment and the growth of the firm. *Economic Journal* 66: 220–235.

Penrose, E.T. 1959. *The theory of the growth of the firm*. Oxford: Blackwell.

Reuber, G.L. 1973. *Private foreign investment in development*. Oxford: Clarendon Press.

Rugman, A.M. 1981. *Inside the multinationals: The economics of internal markets*. New York: Columbia University Press.

Stopford, J.M., and L.T. Wells Jr. 1972. *Managing the multinational enterprise: Organization of the firms and ownership of the subsidiaries*. New York: Basic Books.

Torre, J. de la, 1971. Exports of manufactured goods from developing countries. *Journal of International Business*, Spring.

United Nations, Department of Economic and Social Affairs. 1974. *Multinational corporations in World development*. New York: United Nations.

United Nations, Commission on Transnational Corporations. 1978. Transnational corporations in World development. New York: United Nations. Updated 1982.

Vernon, R. 1971. *Sovereignty at bay: The multinational spread of U.S. enterprises*. New York: Basic Books.

Vernon, R. 1977. *Storm over the multinationals: The real issues*. Cambridge, MA: Harvard University Press.

Warren, B. 1980. *Imperialism: Pioneer of capitalism*. London: New Left Books and Verso.

Wells, L.T. Jr. 1983. *Third world multinationals*. Cambridge, MA: MIT.

Wilkins, M. 1970. *The emergence of multinational enterprise: American business abroad from the colonial Era to 1914*. Cambridge, MA: Harvard University Press.

Wilkins, M. 1974. *The maturing of multinational enterprise: American business abroad from 1914 to 1970*. Cambridge, MA: Harvard University Press.

Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Multiple Equilibria in Macroeconomics

Costas Azariadis

Abstract

The multiple equilibrium literature seeks explanations for excessive economic volatility, persistent poverty, market fads and fashions, and related macroeconomic phenomena that appear to be anomalies in standard models of rational economic behaviour. Terms like *animal spirits, sunspots, irrational exuberance, indeterminacy*, and *bubbles* describe situations of multiple equilibrium. All such ideas assert that future values of macroeconomic states cannot be predicted accurately from knowledge of economic fundamentals. This article describes four types of multiple equilibria common in macroeconomics (missing initial conditions, multiple laws of motion, multiple attractors, and non-fundamental state variables), discusses their causes and reviews what they teach us about economic policy.

Keywords

Animal spirits; Bubbles; Extraneous random variables; Imperfect asset markets; Income effect; Increasing returns; Indeterminacy; Jump variables; Missing initial conditions; Multiple equilibria in macroeconomics; Multiple laws of motion; Overlapping generations models; Public debt; Stable manifolds

JEL Classification

D4; D10

The multiple equilibrium literature seeks explanations for excessive economic volatility, persistent poverty, market fads and fashions, and related macroeconomic phenomena that appear to be anomalies in standard models of rational economic behaviour. Terms like *animal spirits, sunspots, irrational exuberance, indeterminacy*, and *bubbles* describe situations of multiple equilibrium. All of these ideas assert that future values of macroeconomic states cannot be predicted accurately from current values of these states or from knowledge of economic fundamentals, even if households and firms behave with complete rationality.

Most of the economics research community has been sceptical of multiple equilibrium (cf. McCallum 1990), believing that it undermines the comparative statics and comparative dynamics exercises that are essential for policy evaluation and econometric prediction. Is it unreasonable, ask the sceptics, to know how the economy selects one equilibrium when many are possible, and how the expectations of economic actors settle on that particular outcome?

Economists have to weigh these legitimate reservations against direct evidence from laboratory experiments that beliefs do matter (Duffy and Fisher 2005) as well as against the continuing difficulties of unique equilibrium models to come to grips with an expanding array of empirical anomalies in many sub-fields of macroeconomics, from excessively volatile asset prices and exchange rates to persistent underdevelopment. This article describes briefly four types of multiple equilibria common in macroeconomics, discusses what causes them, and reviews briefly what they teach us about economic policy.

M

Typology and Examples

Multiple equilibria occur in dynamic economies whenever the laws of motion that describe macroeconomic states over time admit more than one solution sequence or, more broadly, several asymptotic states. The simplest mathematical example is a set valued, piecewise linear, deterministic law of motion for a scalar state variable $x(t)$, expressed in terms of a vector $v = (A, B, m, a, b)$ of fundamental parameters:

$$\begin{aligned} x(t+1) &= f(x(t), v) = mx(t) + a \quad \text{if} \quad 0 \\ &< x(t) < A = g(x(t), v) \\ &= mx(t) + b \quad \text{if} \quad B < x(t) \end{aligned} \tag{1}$$

for all $t = 0, 1, \ldots$, *with* $0 < m < 1$, $0 < A$, $0 < B$, $0 < a < b$, *and possibly some initial condition* $x(0) > 0$ *fixed by history.*

For different values of the parameter vector v, Eq. (1) illustrates explicitly three major types of multiple equilibria: *indeterminacy from missing initial conditions*, *indeterminacy from multiple laws of motion*, and *multiple attractors*. A fourth type, *non-fundamental state variables* or *sunspots*, occurs when we randomly combine the two laws of motion f and g. All four types are associated with excessively volatile behaviour, that is, with macroeconomic states exhibiting abnormal sensitivity to small changes in fundamentals.

Missing initial conditions is the simplest and best-known type of indeterminacy. Suppose, for example, that there is a unique law of motion f, that is, the parameters A and B are infinitely large. If $x(0)$ is an initial price or, more generally, a *jump variable* that is not predetermined by history but emerges instead from forward-looking markets, then there is a one-dimensional continuum of solutions $x(t,a)$ to Eq. indexed on the indeterminate initial condition $x(0)$:

$$\begin{aligned}\log(x(t,a) - a = (1-m)) \\ = t\log m) + \log(x(0) - a/(1-m))\end{aligned} \tag{2}$$

More generally, an indeterminacy with $S-I$ degrees of freedom appears in any dynamic economy when: (a) history predetermines I initial conditions; (b) the law of motion has S stable eigenvalues; and (c) $I < S$. Equation (2) illustrates the case $(S, I) = (1, 0)$. A major set of economic examples for this kind of multiplicity comes from overlapping generations models. Fiat money in a dynamically inefficient exchange economy (Wallace 1980) has an indeterminate steady state with worthless money at which $(S, I) = (1, 0)$ because history does not fix the initial price of money. Public debt in a dynamically inefficient production economy (Diamond 1965) leads to an indeterminate steady state, with worthless public debt and $(S, I) = (2, 1)$ because the price of debt is also a jump variable. Finally, two-sector growth environments (Galor 1992), in which the distribution of capital between sectors is again a jump variable, exhibit indeterminacy with $(S, I) = (2, 1)$ whenever the consumption good is more capital-intensive than the investment good.

Multiple laws of motion describe a less understood but more pernicious kind of indeterminacy that arises even if there are no jump variables. Examples of this phenomenon are growth models with private information or limited enforcement (Azariadis and Smith 1998; Azariadis and Kaas 2008) as well as Markov switching models in time-series econometrics and empirical finance (Hamilton 1994). To illustrate, let us choose the parameter vector v in Eq. (1) so that

$$(1-m)B < a < b < (1-m)A \tag{3}$$

Then the two laws of motion, f and g, overlap in the interval (B, A); each of them has a steady state, $a = (1-m)$ and $b = (1-m)$ respectively, which is a suitable initial condition for the *other* law. If $x(t,a)$ and $x(t,b)$ are dynamic equilibria for the two laws in Eq. (2), then for any initial condition $x(0)$ in the interval (B, A), we can write down a deterministic general solution $z(t)$ that combines regimes f and g in *any arbitrary time sequence*, that is,

$$\begin{aligned} z(t) &= x(t,a) \quad \text{for some } t \\ &= x(t,b) \quad \text{for all other } t \end{aligned} \tag{4}$$

For each $x(0)$, we may freely select either regime in each time period. In particular, choosing the same regime every period leads to the steady state of that regime; switching regimes periodically leads to deterministic periodic cycles, as in Grandmont (1985), and so on.

Sunspot equilibria are mixtures of multiple deterministic equilibria – static ones as in Cass and Shell (1983) or dynamic ones as in Azariadis (1981) – connected by a *non-fundamental* or *extraneous* random variable. Market sentiment, investor beliefs, and consensus forecasts are three examples of extraneous random variables which often take on more colourful names like 'animal spirits', 'sunspots' or 'self-fulfilling prophecies'. A simple illustration of a non-fundamental state variable is a lottery $s(t)$ played each period over the intercept, a or b, of the two laws of motion in Eq. (1). For instance, if $s(t)$ is a two-state Markov process, then $s(t) = s(t-1)$, with probability $p(a)$ if $s(t-1) = a$, and with probability $p(b)$ if $s(t-1) = b$. The general

stochastic solution $Z(t, s(t))$ to Eq. (1) shows how outcomes depend on the non-fundamental macroeconomic state $s(t)$. Specifically,

$$\begin{aligned} &\text{If } s(t-1)=a, \text{ then } z(t,s(t))= \\ &x(t,a) \ w.p.p(a)=x(t,b) \ w.p.1 \\ &-p(a) \text{ If } s(t-1)=b, \text{ then } z(t,s(t))=x(t,a) \ w. \\ &p.1-p(b)=x(t,b) \ w.p.p(b) \end{aligned} \tag{5}$$

The last type of non-uniqueness, *multiple attractors*, describes environments with several asymptotic states. Here long-run values of macroeconomic states depend on the corresponding initial values, as in Murphy et al. (1989), Azariadis and Drazen (1990), and Matsuyama (1991). We call these environments 'non-ergodic' or ones in which 'history matters'. For example, suppose we pick the parameter vector v in Eq. (1) to eliminate the overlap between regimes f and g, and obtain one piecewise linear law of motion. Specifically, we replace (3) by

$$a < (1-m)A < (1-m)B < b \tag{6}$$

Then, for each initial $x(0)$, the general deterministic solution $z(t)$ to Eq. (1) is a unique step function, which traces the law f up to $x = A$, and jumps to the other law g at that point. Mathematically,

$$\begin{aligned} z(t) &= x(t,a) \ \textit{if} \ z(t-1) < A \\ &= x(t,b) \ \textit{if} \ z(t-1) > A \end{aligned} \tag{7}$$

Equilibrium here is completely determinate and utterly predictable if history fixes $x(0)$, but the asymptotic state is $a = (1-m)$ if $x(0) < A$, and $b = (1-m)$ if $x(0) > A$. History *matters* in this situation because small or temporary shocks to the macroeconomic state $z(t)$ can have substantial and long-lasting consequences if that state is anywhere near the critical value A.

Causes

Dynamic inefficiency and dynamic complementarities are the two most common proximate causes of multiple equilibrium in macroeconomic models. Dynamic inefficiency is a property of economies with very patient consumers who are energetic savers at low interest rates. For example, holders of short-term US Treasury bills in the last 50 years seem content with an average real pre-tax annual yield of about 1%. Very patient savers are willing to invest in *bubbles*, paying top dollar for assets with low dividends. Bubbles themselves (Tirole 1985; Shiller 1989) are notoriously indeterminate objects in their initial conditions and laws of motion; they may deflate now, later or not at all, depending on investor sentiment.

Economies with externalities, increasing returns and, most notably, imperfect asset markets often exhibit complementarities in production or consumption which cause excess demands for consumption goods and productive factors to bend backward instead of sloping downward. The typical outcome is several steady states and several laws of motion or *stable manifolds*, each one leading to a distinct asymptotic state. In particular, multiple equilibria occur when externalities or increasing returns link the payoffs of each agent with the actions of others, both in strategic environments (Cooper and John 1988) and in competitive ones (Benhabib and Farmer 1994). Producers, for example, find it advantageous to raise, hold steady, or lower output in tandem with their industry or the whole national economy.

Imperfect asset markets, especially restrictions on debt and short sales (Bewley 1986; Kehoe and Levine 1993; Kiyotaki and Moore 1997) are an intellectually bountiful and empirically compelling source of complementarities in consumption. This literature motivates restrictions on short sales by the collateral requirements of creditors and, more generally, as a deterrent to debtor default. Short-sales constraints depend on the excess payoff of solvency (which guarantees unfettered participation in future asset markets) over default (which restricts trading in future asset markets). Constraints on short sales are tighter the smaller this excess payoff is because smaller excess payoffs strengthen the temptation to default.

Debt constraints cause two dynamic complementarities in consumption, one through prices and the other through quantities

(Azariadis and Kaas 2007). Either one may be sufficient to overcome the intertemporal substitution effect embedded in the consumer's utility function. Specifically, price changes create a dynamic complementarity when the ordinary income effect is amplified by a relaxation of binding short-sale restrictions. The same outcome is achieved by quantity changes when an anticipated relaxation of future constraints increases the current payoff to solvency, and to continued market participation, thus slackening today's constraints.

Lessons for Policy

What is the function of economic policy in a deterministic world of many steady states like the one described in Eq. (7)? What should policy do in the stochastic world of Eq. (5) where non-fundamental variables like beliefs, forecasts, consumer sentiment, 'sunspots', or 'animal spirits' could be every bit as important as fundamentals? Dynamic economies with several asymptotic states have two special properties: long-run performance depends on the starting state $x(0)$; and temporary shocks may have permanent consequences. Any economy that is headed towards an inferior or undesirable steady state may be shocked temporarily until it finds a path leading to a more desirable state. In growth models with many asymptotic states, these shocks are easy to achieve in principle via short-lasting gifts of physical or human capital, by forgiving international debt, and so on. The US-supported Marshall Plan for Europe did exactly that in the 1940s and 1950s. Africa seems in need of a similar plan now but the internal situation in that continent is more problematic than Europe's was at the end of the Second World War.

A bigger conceptual, as distinct from political, challenge is to formulate policies appropriate for environments swayed by non-fundamental variables and vulnerable to spurious volatility. If equilibria were well described by the stochastic process of Eq. (5), could we find an economic policy to eliminate the unnecessary randomness, and bolster among consumers the belief that the economy is headed toward the more desirable of the two steady states, say, $b/(1-m)$? Viewing economic policy as *equilibrium selection* is fairly widespread in the monetary policy literature (Woodford 2003), and broadly consistent with monetary neutrality. On this view, credible monetary policy may be unable to influence the set of possible long-run equilibria, but it does bear on which one the economy selects. In Eq. (5), for example, reactive policy rules may be unable to change the laws of motion f and g but they can still deliver the long-run state $b/(1-m)$ if they influence the public's beliefs about the long-run likelihood of each state. All it takes to achieve the high state is nudging the two mixing probabilities, $p(a)$ towards zero and $p(b)$ towards 1.

See Also

- ▶ Animal Spirits
- ▶ Bubbles

Bibliography

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25: 380–396.

Azariadis, C., and A. Drazen. 1990. Threshold externalities in economic development. *Quarterly Journal of Economics* 105: 501–526.

Azariadis, C., and L. Kaas. 2008. Credit and growth under limited commitment. *Macroeconomic Dynamics* 12(Supp. 1): 20–30.

Azariadis, C., and B. Smith. 1998. Financial intermediation and regime switching in business cycles. *American Economic Review* 88: 516–536.

Benhabib, J., and R. Farmer. 1994. Indeterminacy and increasing returns. *Journal of Economic Theory* 63: 19–41.

Bewley, T. 1986. Dynamic implications of the form of the budget constraint. In *Models of economic dynamics*, ed. H. Sonnenschein. New York: Springer.

Cass, D., and K. Shell. 1983. Do sunspots matter? *Journal of Political Economy* 91: 193–227.

Cooper, R., and A. John. 1988. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics* 103: 441–464.

Diamond, P. 1965. National debt in a neoclassical growth model. *American Economic Review* 55: 1126–1150.

Duffy, J., and E. Fisher. 2005. Sunspots in the laboratory. *American Economic Review* 95: 510–529.

Galor, O. 1992. A two-sector overlapping generations model. *Econometrica* 60: 351–386.

Grandmont, J.-M. 1985. On endogenous competitive business cycles. *Econometrica* 53: 995–1045.
Hamilton, J. 1994. *Time series analysis*. Princeton: Princeton University Press.
Kehoe, T., and D. Levine. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60: 865–888.
Kiyotaki, N., and J. Moore. 1997. Credit cycles. *Journal of Political Economy* 105: 221–248.
Matsuyama, K. 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. *Quarterly Journal of Economics* 106: 617–650.
McCallum, B. 1990. New classical macroeconomics: A sympathetic account. In *The state of macroeconomics*, ed. S. Honkapohja. Oxford: Basil Blackwell.
Murphy, K., A. Shleifer, and R. Vishny. 1989. Industrialization and the big push. *Journal of Political Economy* 97: 1003–1026.
Shiller, R. 1989. *Market volatility*. Cambridge, MA: MIT Press.
Tirole, J. 1985. Asset bubbles and overlapping generations. *Econometrica* 53: 1499–1528.
Wallace, N. 1980. The overlapping generations model of fiat money. In *Models of monetary economies*, ed. J. Kareken and N. Wallace. Minneapolis: Federal Reserve Bank of Minneapolis.
Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Multiple Testing

Joseph P. Romano, Azeem M. Shaikh and Michael Wolf

Abstract

Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis. If decisions about the individual hypotheses are based on the unadjusted marginal p-values, then there is typically a large probability that some of the true null hypotheses will be rejected. Unfortunately, such a course of action is still common. In this article, we describe the problem of multiple testing more formally and discuss methods which account for the multiplicity issue. In particular, recent developments based on resampling result in an improved ability to reject false hypotheses compared to classical methods such as Bonferroni.

Keywords

Multiple testing; Familywise error rate; Real estate finance; Resampling

JEL Classifications

c12

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. This scenario is quite common in much empirical research in economics. Some examples include: (i) one fits a multiple regression model and wishes to decide which coefficients are different from zero; (ii) one compares several forecasting strategies to a benchmark and wishes to decide which strategies are outperforming the benchmark; (iii) one evaluates a program with respect to multiple outcomes and wishes to decide for which outcomes the program yields significant effects.

If one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected by chance alone may be unduly large. Take the case of $S = 100$ hypotheses being tested at the same time, all of them being true, with the size and level of each test exactly equal to α. For $\alpha = 0.05$, one expects five true hypotheses to be rejected. Further, if all tests are mutually independent, then the probability that at least one true null hypothesis will be rejected is given by $1–0.95^{100} = 0.994$.

Of course, there is no problem if one focuses on a particular hypothesis, and only one of them, a priori. The decision can still be based on the corresponding marginal p-value. The problem only arises if one searches the list of p-values for significant results a posteriori. Unfortunately, the latter case is much more common.

Notation

Suppose data X is generated from some unknown probability distribution P. In anticipation of asymptotic results, we may write $X = X^{(n)}$, where n typically refers to the sample size.

A model assumes that P belongs to a certain family of probability distributions, though we make no rigid requirements for this family; it may be a parametric, semiparametric, or nonparametric model.

Consider the problem of simultaneously testing a hypothesis H_s against the alternative hypothesis H_s'' for $s = 1, \ldots, S$. A multiple testing procedure (MTP) is a rule which makes some decision about each H_s. The term *false discovery* refers to the rejection of a true null hypothesis. Also, let $I(P)$ denote the set of true null hypotheses, that is, $s \in I(P)$ if and only if (iff) H_s is true.

We also assume that a test of the individual hypothesis H_s is based on a test statistic $T_{n,s}$, with large values indicating evidence against H_s. A marginal p-value for testing H_s is denoted by $\widehat{p}_{n,s}$.

Familywise Error Rate

Accounting for the multiplicity of individual tests can be achieved by controlling an appropriate *error rate*. The traditional or classical *familywise error rate* (FWE) is the probability of one or more false discoveries:

$$FWE_P = \{\text{reject at least one hypothesis } H_s : s \in I(P)\}.$$

Control of the FWE means that, for a given significance level α,

$$\text{FWE}_P \leq \text{A for any } P. \tag{1}$$

Control of the FWE allows one to be $1 - \alpha$ confident that there are no false discoveries among the rejected hypotheses.

Note that 'control' of the FWE is equated with 'finite-sample' control: (1) is required to hold for any given sample size n. However, such a requirement can often only be achieved under strict parametric assumptions or for special permutation setups. Instead, we then settle for *asymptotic* control of the FWE:

$$\limsup_{n\to\infty} \text{FWE}_P \leq \text{A for any } P. \tag{2}$$

Methods Based on Marginal *p*-values

MTPs falling in this category are derived from the marginal or individual p-values. They do not attempt to incorporate any information about the dependence structure between these p-values. There are two advantages to such methods. First, we might only have access to the list of p-values from a past study, but not to the underlying complete data set. Second, such methods can be very quickly implemented. On the other hand, as discussed later, such methods are generally sub-optimal in terms of power.

To show that such methods control the desired error rate, we need a condition on the p-values corresponding to the true null hypotheses:

$$H_s \text{ true} \Leftrightarrow s \in I(P) \rightarrow P\{\widehat{p}_{n,s} \leq u\} \leq u \text{ for any } u \in (0,1). \tag{3}$$

Condition (3) merely asserts that, when testing H_s alone, the test that rejects H_s when $\widehat{p}_{n,s} \leq u$ has level u, that is, $\widehat{p}_{n,s}$ is a proper p-value.

The classical method to control the FWE is the Bonferroni method, which rejects H_s iff $\widehat{p}_{n,s} \leq A/S$: More generally, the weighted Bonferroni method rejects H_s if $\widehat{p}_{n,s} \leq w_s \cdot A/S$; where the constants w_s, satisfying $w_s \geq 0$ and $\Sigma_s w_s = 1$, reflect the 'importance' of the individual hypotheses.

An improvement is obtained by the method of Holm (1979). The marginal $p$values are ordered from smallest to largest: $\widehat{p}_{n,(1)} \leq \widehat{p}_{n,(2)} \leq \ldots \leq \widehat{p}_{n,(S)}$ with their corresponding null hypotheses labeled accordingly: $H_{(1)}, H_{(2)}, \ldots, H_{(s)}$. Then, $H_{(s)}$ is rejected iff $\widehat{p}_{n,(j)} \leq A/(S - j + 1)$ for $j = 1, \ldots, s$. In other words, the method starts with testing the most significant hypothesis by comparing its p-value to α/S, just as the Bonferroni method. If the hypothesis is rejected, the method moves on to the second most significant hypothesis by comparing its p-value to $\alpha/(S-1)$, and so on, until the procedure comes to a stop. Necessarily, all hypotheses rejected by Bonferroni will also be rejected by Holm, but potentially a few more will be too. So, trivially, the method is more powerful. But it still controls the FWE under (3).

If it is known that the p-values are suitably positive dependent, then further improvements can be obtained with the use of Simes identity; see Sarkar (1998).

So far, we have assumed 'finite-sample validity' of the null p-values expressed by (3). However, often p-values are derived by asymptotic approximations or resampling methods, only guaranteeing 'asymptotic validity' instead:

$$H_s \text{ true} \Leftrightarrow s \in I(P) \rightarrow \limsup_{n\to\infty} P\left\{\widehat{p}_{n,s} \leq u\right\} \leq u \text{ for any } u \in (0,1). \quad (4)$$

Under this more realistic condition, the MTPs presented in this section only provide asymptotic control of the FWE in the sense of (2).

Single-step Versus Stepwise Methods

In single-step MTPs, individual test statistics are compared to their critical values simultaneously, and after this simultaneous 'joint' comparison, the multiple testing method stops. Often there is only one common critical value, but this need not be the case. More generally, the critical value for the sth test statistic may depend on s. An example is the weighted Bonferroni method discussed above.

Often single-step methods can be improved in terms of power via stepwise methods, while still maintaining control of the desired error rate. Stepdown methods start with a single-step method but then continue by possibly rejecting further hypotheses in subsequent steps. This is achieved by decreasing the critical values for the remaining hypotheses depending on the hypotheses already rejected in previous steps. As soon as no further hypotheses are rejected, the method stops. The Holm (1979) method discussed above is a stepdown method.

Stepdown methods therefore improve upon single-step methods by possibly rejecting 'less significant' hypotheses in subsequent steps. In contrast, there also exist stepup methods that start with the least significant hypotheses, having the smallest test statistics, and then 'step up' to further examine the remaining hypotheses having larger test statistics.

More general methods that control the FWE can be obtained by the closure method; see Hochberg and Tamhane (1987).

Resampling Methods Accounting for Dependence

Methods based on p-values often achieve (asymptotic) control of the FWE by assuming (i) a worst-case dependence structure or (ii) a 'convenient' dependence structure (such as mutual independence). This has two potential disadvantages. In case of (i), the method can be quite sub-optimal in terms of power if the true dependence structure is quite far away from the worst-case scenario. In case of (ii), if the convenient dependence structure does not hold, even asymptotic control may not result. As an example for case (i), consider the Bonferroni method. If the p-values were perfectly dependent, then the cut-off value could be changed from α/S to α. While perfect dependence is rare, this example serves to make a point. In the realistic scenario of 'strong cross-dependence', the cut-off value could be changed to something a lot larger than α/S while still maintaining control of the FWE. Hence, it is desirable to account for the underlying dependence structure.

Of course, this dependence structure is unknown and must be (implicitly) estimated from the available data. Consistent estimation, in general, requires that the sample size grow to infinity. Therefore, in this subsection, we will settle for asymptotic control of the FWE. In addition, we will specialize to making simultaneous inference on the elements of a parameter vector $\theta = (\theta_1, \ldots, \theta_S)^T$. Assume the individual hypotheses are one-sided of the form:

$$H_s : \Theta_s \leq 0 \text{ vs. } H_s'' : \Theta_s > 0. \quad (5)$$

Modifications for two-sided hypotheses are straightforward.

The test statistics are of the form $T_{n,s} = \widehat{\Theta}_{n,s}/\widehat{\Sigma}_{n,s}$. Here, $\widehat{\Theta}_{n,s}$ is an estimator of θ_s computed

from $X_{(n)}$. Further, $\widehat{\Sigma}_{n,s}$ is either a standard error for $\widehat{\Sigma}_{n,s}$ or simply equal to $1/\sqrt{n}$ in case such a standard error is not available or only very difficult to obtain.

We start by discussing a single-step method. An idealized method would reject all H_s for which $T_{n,s} \geq d_1$ where d_1 is the $1-\alpha$ quantile under P of the random variable $max_s\left(\widehat{\Theta}_{n,s} - \Theta_s\right)/\widehat{\Sigma}_{n,s}$ Naturally, the quantile d_1 does not only depend on the marginal distributions of the centered statistics $\left(\widehat{\Theta}_{n,s} - \Theta_s\right)/\widehat{\Sigma}_{n,s}$ but, crucially, also on their dependence structure.

Since P is unknown, the idealized critical value d_1 is not available. But it can be estimated consistently under weak regularity conditions as follows. Take $\widehat{d}_1$ as the $1-\alpha$ quantile under $\widehat{P}_n$ of $\max_s\left(\widehat{\Theta}^*_{n,s} - \widehat{\Theta}_{n,s}\right)/\widehat{\Sigma}^*_{n,s}$. Here, $\widehat{P}_n$ is an *unrestricted* estimate of P. Further $\widehat{\Theta}^*_{n,s}$ and $\widehat{\Sigma}^*_{n,s}$ are the estimator of θ_s and its standard error (or simply $1/\sqrt{n}$), respectively, computed from $X^{(n),*}$ where $X^{(n),*} \sim \widehat{P}_n$. In other words, we use the bootstrap to estimate d_1. The particular choice of $\widehat{P}_n$ depends on the situation. In particular, if the data are collected over time a suitable time series bootstrap needs to be employed; see Davison and Hinkley (1997) and Lahiri (2003).

We have thus described a single-step MTP. However, a stepdown improvement is possible. In any given step j, we simply discard the hypotheses that have been rejected so far and apply the single-step MTP to the remaining universe of non-rejected hypotheses. The resulting critical value $\widehat{d}_j$ necessarily satisfies $\widehat{d}_j \leq \widehat{d}_{j-1}$ so that new rejections may result; otherwise the method stops.

This bootstrap stepdown MTP provides asymptotic control of the FWE under remarkably weak regularity conditions. Mainly, it is assumed that $\sqrt{n}\left(\widehat{\Theta} - \Theta\right)$ converges in distribution to a (multivariate) continuous limit distribution and that the bootstrap consistently estimates this limit distribution. In addition, if standard errors are employed for $\widehat{\Sigma}_{n,s}$, as opposed to simply using $1/\sqrt{n}$, it is assumed that they converge to the same non-zero limiting values in probability, both in the 'real world' and in the 'bootstrap world'. Under even weaker regularity conditions, a subsampling approach could be used instead; see Romano and Wolf (2005). Furthermore, when a randomization setup applies, randomization methods can be used as an alternative; see Romano and Wolf (2005) again.

Related methods are developed in White (2000) and Hansen (2005). However, both works are restricted to single-step methods. In addition, White (2000) does not consider studentized test statistics. Stepwise bootstrap methods to control the FWE are already proposed in Westfall and Young (1993). An important difference in their approach is that they bootstrap under the joint null, that is, they use a *restricted* estimate of P where the contraints of all null hypotheses jointly hold. This approach requires the so-called *subset pivotality* condition and is generally less valid than the approaches discussed so far based on an unrestricted estimate of P; for instance, see Example 4.1 of Romano and Wolf (2005).

Generalized Error Rates

So far, attention has been restricted to the FWE. Of course, this criterion is very strict; not even a single true hypothesis is allowed to be rejected. When S is very large, the corresponding multiple testing procedure (MTP) might result in low power, where we loosely define 'power' as the ability to reject false null hypotheses.

Let F denote the number of false rejections and let R denote the total number of rejections. *The false discovery proportion* (FDP) is defined as $\text{FDP} = (F/R) \cdot 1\{R > 0\}$, where $1\{\cdot\}$ denotes the indicator function. Instead of the FWE, we may consider the probability of the FDP exceeding a small, pre-specified proportion: $P\{\text{FDP} > \gamma\}$, for some $\gamma \in [0,1)$. The special choice of $\gamma = 0$ simplifies to the traditional FWE. Another alternative to the FWE is the *false discovery rate* (FDR), defined to be the expected value of the FDP: $\text{FDR}_\text{P} = E_\text{P}$ (FDP).

By allowing for a small (expected) fraction of false discoveries, one can generally gain a lot of power compared with FWE control, especially when S is large. For the discussion of MTPs to

provide (asymptotic) control of the FDP and the FDR, the reader is referred to Romano et al. (2008a, b) and the references therein.

Bibliography

Davison, A.C., and D.V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge: University Press, Cambridge.

Hansen, P.R. 2005. A test for superior predictive ability. *Journal of Business and Economics Statistics* 23: 365–80.

Hochberg, Y., and A. Tamhane. 1987. *Multiple comparison procedures*. New York: Wiley.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65–70.

Lahiri, S.N. 2003. *Resampling methods for dependent data*. New York: Springer.

Romano, J.P., and M. Wolf. 2005. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* 100: 94–108.

Romano, J.P., A.M. Shaikh, and M. Wolf. 2008a. Control of the false discovery rate under dependence using the bootstrap and subsampling (with discussion). *Test* 17: 417–42.

Romano, J.P., A.M. Shaikh, and M. Wolf. 2008b. Formalized data snooping based on generalized error rates. *Econometric Theory* 24: 404–47.

Sarkar, S.K. 1998. Some probability inequalities for ordered M T P2 random variables: A proof of simes conjecture. *Annals of Statistics* 26: 494–504.

Westfall, P.H., and S.S. Young. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.

White, H.L. 2000. A reality check for data snooping. *Econometrica* 68: 1097–126.

Multiplier Analysis

Edward J. Amadeo

Keywords

Equilibrium; Excess demand; Expectations; Hicks, J. R.; Kalecki, M.; Keynes, J. M.; Investment; Investment multiplier; Multiplier analysis; Real wage rate; Saving; Saving equals investment; Shifting equilibrium; Wicksell, J. G. K.

JEL Classifications

E2

What is the effect of a change in the level of investment? Wicksell (1935) was the first economist to pose this question explicitly in the context of his 'pure credit economy'. Voluntary or anticipated saving is not a requirement if the banking system is willing to supply the necessary credit to finance an increase of investment demand. The effect of this increase of investment demand is an increase in the level of prices (if the level of output is fixed or given), or output if there is idle capacity and unemployed labour.

In his *Treatise on Money* (1930) Keynes analyses the same question. Just as in Wicksell's model, in the *Treatise*, investment is independent of current saving. The effects of a change of investment are studied through the *Treatise*'s 'Fundamental Equations' according to which a difference between current (or voluntary) saving and investment will give rise to a change in the price level. It is a pure excess demand effect. Changes in the price level will lead to unforeseen (or windfall) profits or losses which, in turn, will affect producers' next period decision to produce and employ. Windfall profits will have the effect of inducing producers to increase the level of output; losses will have the opposite effect. The effect may not be as mechanical as described here if new informations (concerning, for example, changes in economic policies) come into the picture.

Book IV of the *Treatise* studies the 'credit cycle', that is, the effects of changes in monetary or banking policies on the rate of interest which may have an effect on the decisions to save and invest, and therefore, on the price and output levels.

Changes in both the price and output levels are seen as deviations from their long- period or equilibrium counterparts; they are short-period or disequilibrium levels of price and output which, so to speak, oscillate around the equilibrium as defined by the equality between voluntary saving and investment. However, just as in Wicksell's analysis, once the system deviates from the equilibrium

position, very little is said in terms of the path towards a new equilibrium; indeed, the latter is not really determined.

Multiplier analysis is very much related to the adjustment process described above. The real *differentia* is that it focuses predominantly on the notions of stability and equilibrium of the process. The most important contributors for the development of the multiplier analysis were Kahn (1931), Keynes (1936) and Kalecki (1971).

The Multiplier as an Exercise in Comparative Statics

Let us consider the effects of a change in the level of investment which is known to all the relevant agents of the economy. Also let us temporarily assume that producers of consumption goods fully anticipate the effects of this change in investment on the demand for their products. An increase in the level of investment demand implies a greater level of production of capital goods. The degrees of capacity utilization and employment in the capital goods sector increase, thus leading to higher profits and a greater wage bill. Part of the extra profits and wages earned will be spent in consumption goods; the rest will be saved. The share of profits and wages spent in consumption goods are determined respectively by the propensities to consume out of profits and wages. These, according to Keynes (1936, chapters 8 and 9), depend on objective factors (other than income) such as the money wage rate and agents' rates of time-discounting, and subjective factors such as precaution and avarice.

Thus the main effect of an increase in investment is that it induces an increase in consumption, saving, and income. The final effect on the level of income will depend essentially on the propensity to consume of the economy. The greater the propensity to consume, the greater will be the increase in the demand for consumption goods resulting from an initial increase in the income generated in the capital goods sector. The immediate effect on the demand for consumption goods will be given by $C = cI$ where C and I are respectively the levels of consumption and investment, and c is the weighted average of the propensities to consume out of profits and wages. The immediate effect on the level of income will be given by $\Delta Y = \Delta I + c\Delta I$. Note that a second round of the multiplier process will lead to an increase in the level of income given by $\Delta Y = \Delta I + c\Delta I + c^2\Delta I$. In the limit the effect will be given by $\Delta Y = \Delta I + c\Delta I + c^2\Delta I + \ldots = [1/(1 - c)]\Delta I$. The term $1/(1 - c)$ is called the investment multiplier. According to Keynes, the multiplier 'tells us that, when there is an increment of aggregate investment, income will increase by an amount which is $[1/(1 - c)]$ times the increment in investment' (Keynes 1936, p. 115).

Note that the change in the level of saving (ΔS) is given by the propensity to save ($s = 1 - c$) times the level of income, that is, $\Delta S = s\Delta Y$, which, according to the above analysis, is also equal to the initial change in the level of investment. Thus, through the multiplier mechanism, a change in the level of investment gives rise to an equal level of saving. The multiplier is essentially an equilibrating mechanism. It refers to the adjustment of the economy given an exogenous change, and it determines the equilibrium levels of income and saving associated with different levels of investment demand. It describes the changes in the level of consumption which eventually makes the latter compatible to each level of investment given the propensity to consume of the economy.

The essential difference between the multiplier mechanism and the description of credit cycles found in Keynes's *Treatise on Money* as well as in the analyses of Wicksell and the Swedish economists (Ohlin and Lindhal for example), is that it emphasizes the notion of equilibrium. It determines the new equilibrium configuration associated with any change in the level of investment demand rather than only its immediate effects. Because it is an equilibrating mechanism it must also take into account the stability conditions of the process. In terms of the simple static version discussed above, the only stability condition is that the propensity to consume must be smaller than one. If it were greater than one the system would always explode either to a situation of full employment or zero-employment of the labour force and capacity utilization. As noted by Keynes, 'if the [community] seek to consume

the whole of any increment in income, there will be no point of stability and prices with rise without limit' (Keynes 1936, p. 117). However, since the propensity to consume is always positive, the multiplier is always greater than one which implies that fluctuations in investment will lead to fluctuations of income of greater magnitude. Thus, the workings of the multiplier mechanism itself may be regarded as a source of instability.

The Multiplier as an Exercise in Dynamics

What makes the analysis of the above section static is the fact that it emphasizes the equilibrium configuration associated with a given (and known) level of investment, and a given propensity to consume. The decision to consumer is rather passive and taking it into account does not really make the analysis dynamic. What is most important, however, is that the decisions to produce are not considered. Production takes time, and therefore decisions to produce involve expectations over a period of time. A dynamic approach to the analysis of the multiplier should emphasize the role of time and expectations associated with the decisions to produce.

What is the appropriate time unit for the analysis of the multiplier process if decisions to produce are to be considered explicitly? Following Keynes we shall take the short period as the appropriate time unit. The short period is associated with 'daily' decisions, and daily here stands 'for the shortest interval of time after which the firm is free to revise its decisions as to how much employment to offer' (Keynes 1936, p. 47). Producers make their decisions as to how much to produce based on their short-period expectations.

On the demand side the object of such expectations are either the expected sale- proceeds or the expected price, that is, the price which the producer expects to get for his product at the end of the period of production. Let us take the expected price as the relevant variable, and assume that the producer knows the remuneration rates of the variable inputs and the shape of his cost curve. Given this information we may assume that the producer goes through the following optimization exercise in order to determine the levels of output and employment: max $E[p]X - wN$ *st* . $X = F(N, K)$ where $E[p]$ is the expected price, X and N are the levels of output and employment respectively, w is the money-wage rate, K is the stock of capital (assumed to be given), and F is a production function. The level of employment associated with the expected price must satisfy the following condition: $w/\mathrm{E}[p] = F'(N^*)$. The level of output is obviously $X^* = F(N^*)$.

Let us assume that the level of investment has been stable for a rather long period of time. Producers of consumption goods know not only the level of investment but also the demand for their products associated with this level.

Therefore they are able to form correct expectations concerning the demand for their products, and their price. In short, in each and every period the expected price corresponds to the market price, that is, $E[p] = p$. We now let the level of investment increase but assume that the producers of consumption goods either do not know that the change has taken place or the effect of the change on the demand for their products. If the latter is the case, assume that they underestimate the effect on demand. In either case the actual price will be greater than the expected price associated with the predetermined level of output (X^*), that is, $p > E[p]$ where p is the market price. In this example producers will experience a windfall profit given by $Q = (p - E[p])X^*$. The same exercise could be carried on taking stocks rather than the price as the adjustment variable (see Hicks 1974, chapter 1).

The process initiated with a change in investment demand could go on for a long period. Producers would continue to get their expectations wrong, profits or losses would appear, new decisions would be taken and so on. Will producers ever get their prices (and production decisions) right? If we assume that the level of investment will not be affected by changes in short-period expectations, and depending on the way producers form their expectations, they will eventually converge to an equilibrium position. If, for

example, producers form their short- period expectations in an adaptive fashion, for certain values of the parameters of the expectation function, the system will converge to a position of rest. For other values of the parameters the system will not converge. This only implies that the way producers form their expectations may affect the stability of the multiplier process and the trajectory of the relevant variables.

Does the way producers form their expectations affect the equilibrium configuration? The answer here is no. If the level of investment is assumed to be given and the process is assumed to be stable (which, again, depends on the parameters of the expectation function), the equilibrium configuration will be exactly the same as the one associated with a process in which producers form their expectations in a rational fashion. By 'rational' here we mean that expectations are recurrently correct. Keynes was aware of this result: in his lecture notes written in 1937 he argued that his principle of effective demand is substantially the same independently of the way expectations are formed (see Keynes 1973, pp. 180–1).

The Multiplier and the Notion of 'Shifting Equilibrium'

So far we have examined the multiplier mechanism assuming that either the level or the expected level of investment is given. In both the static and dynamic analyses the multiplier tells us the levels of income and saving compatible with a given level or expected level of investment. The advantage of these approaches to the multiplier is that they emphasize the notion of equilibrium, that is, they provide a definite result to the effect of a change in investment.

However, once the notion of equilibrium has become clear, we should turn our attention to the interactive relation between the level of investment and the workings of the multiplier. The level of investment is quite a volatile variable. Long-period expectations (which play a central role in the determination of the level of investment) change for various reasons. They change due to changes in the political or international environments; due to changes in economic policies; or due to objective problems of individual industries which tend to affect the expected performance of other industries of the economy. To different states of long-period expectations there corresponds different levels of investment and, therefore, different 'levels of long-period employment' (Keynes 1936, p. 48). The extent to which short-period expectations are fulfilled may also affect the level of investment. If the actual demand is persistently greater than the expected demand, producers will tend to revise their long-period expectations and investment decisions.

We may associate the notion of 'shifting equilibrium' with the evolution of the economic system as determined by different states of long-period expectations, and therefore, characterized by a sequence of equilibrium configurations of income and saving. By shifting equilibrium Keynes meant 'the theory of a system in which changing views about the future are capable of influencing the present situation' (1936, p. 293).

Distribution and the Multiplier

The relationship between the distribution of income (or the real wage) and the multiplier depends on assumptions about the exogeneity or endogeneity of the real wage. In the *General Theory* Keynes assumed perfect competition *cum* profit maximization and decreasing marginal returns which, for a given money-wage rate, implies that the real wage is endogenously determined. It also implies that the greater the levels of employment and output, the smaller the real wage. This result has an important implication for the workings of the Keynesian multiplier. If we assume – as Keynes and Kalecki usually do – that the propensity to consume out of wages is greater than the propensity to consume out of other types of incomes (profits, interests, and so on), as the level of income increases and the real

wage falls, the value of the multiplier decreases. Keynes pointed out to this result in the *General Theory*:

> the increase of employment will tend, owing to the effect of diminishing returns, ... to increase the proportion of aggregate income which accrues to the entrepreneurs, whose ... propensity to consume is probably less than the average for the community as a whole. (1936, p. 121.

Kalecki (1971) assumed constant marginal returns and gave up profit maximization. Instead he assumed that firms determine their prices through a markup over variable costs which, in a closed economy, also determines the real wage. Therefore, according to Kalecki, the real wage is exogenously determined, and does not change as the levels of output and employment change. This means that the multiplier does not change either as the level of output changes; it depends on the propensity to consume out of wages and profits and the level of the markup, both assumed to be constant over the cycle.

See Also

► Multiplier–Accelerator Interaction

Bibliography

Hicks, J.R. 1974. *The crisis in Keynesian economics*. Oxford: Blackwell.

Kahn, R. 1931. The relation of home investment to unemployment. *Economic Journal* 41 (June): 173–198. Reprinted in R. Kahn, *Selected essays on employment and growth*. Cambridge: Cambridge University Press, 1972.

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy*. Cambridge: Cambridge University Press.

Keynes, J.M. 1930. *A treatise on money. Vol. 1: The Pure Theory of Money*. London: Macmillan.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Keynes, J.M. 1973. In *The collected writings of John Maynard Keynes*, ed. D.E. Moggridge and E. Johnson, vol. 14. London: Macmillan for the Royal Economic Society.

Wicksell, K. 1935. *Lectures in political economy*. Vol. 2. London: Routledge & Kegan Paul.

Multiplier–Accelerator Interaction

A. Medio

JEL Classifications
E2

The phrase 'multiplier–accelerator' refers to a combination of a theory of income as determined by investment and a theory of investment as determined by the rate of change of income.

The concept of multiplier is usually attributed to Richard Kahn (1931), from whom it was adopted by Keynes and used as a building block for his General Theory. The idea was probably shared by a number of European economists in the 1930s and was certainly known to Michael Kalecki, independently of Keynesian influence.

The theory of multiplier in its pure (and static) form can be described thus. In a capitalist economy, investment can always be realized in real terms. The necessary saving will be made available by means of corresponding variations of the level of income, given the propensity to save. With generally underutilized capacity and labour and fixed prices – the most common hypothesis – *real* income will take whatever value generates a flow of saving equal to planned investment. Alternatively, in the presence of supply constraints, the level of prices will adjust and deflate consumption expenditure so as to make available the real resources required for investment.

The former, 'fixprice' version of this simple relation can be stated in the form of algebraic equations, as follows

$$Y = C + I; \tag{1}$$

$$C = cY, c = 1 - s \tag{2}$$

$$I = \overline{I} \tag{3}$$

where *Y, C, I* indicate, respectively, actual income, consumption and investment; $\overline{I}$ is desired

investment; c and s are the propensities to consume and to save, respectively.

Elementary manipulation yields:

$$Y = (1/s)\bar{I} \tag{4}$$

where (1/s) measures the multiplier and the causal relation runs from right to left.

The concept of accelerator appeared in the economic literature much earlier than the *Generai Theory* and was perhaps first developed by Aftalion (1909) and J.M. Clark (1917). It is based on the idea that the relation between productive capacity (somewhat measured by a scalar quantity, the capital stock) and production can vary only within narrow limits and, in a first approximation, may be taken as a constant.

The constancy of the capital-output ratio may be defended on the basis of two main arguments:

(i) Technical coefficients are fixed (or change little) even though the interest rate may vary: in economists' parlance, the isoquants are L-shaped. Whatever the plausibility of this hypothesis may be from an engineering point of view, it is difficult to accept it on economic grounds. Indeed, when 'capital' is a vectorial quantity (i.e. a list of different capital goods), the capital–output ratio depends both on technical coefficients and on relative prices and the rate of interest.

(ii) Technical coefficients vary, within a certain technology, as functions of the rate of interest. If the latter is constant so are the former.

The assumption on (ii) may be accepted or rejected for lack of realism but is formally correct. On the other hand, it is also consistent with the fix-price approach to income determination. In its starkest form, the accelerator (Harrod called it 'the Relation') can be described by the equation

$$K = vY, \tag{5}$$

(where K indicates the capital stock and v the desired capital–output ratio) or, in its incremental form

$$I = v\dot{Y} \tag{6}$$

where an overdot indicates the derivative with respect to time.

The idea came naturally to combine multiplier and accelerator and derive a model 'complete' in the sense that, given initial conditions, it determines the time evolution of capital stock and income. This was first attempted in the late 1930s by Harrod (1936, 1939) and, in a more mathematical manner, by Samuelson (1939). In the subsequent years, a substantial part of the literature on cycle and growth was also based on the interaction between multiplier and accelerator.

In order to discuss this idea formally, let us couple Eqs. (4) and (6). We shall obtain

$$sY = v\dot{Y} \tag{7}$$

and

$$\left(\dot{Y}/Y\right) = (s/v) \tag{8}$$

Equation (8) represents the proportional rate of growth of income as a function of the propensity to save and the acceleration coefficient and was first investigated by Harrod and Domar, after whom it has been named ever since.

The model described by Eqs. (1, 2, 3, 4, 5, 6, 7 and 8) implicitly assumes that equality always holds between demand (= consumption + investment) and supply (= income), as well as between actual and desired consumptions, the results may become drastically different. This line of research was pursued early by Samuelson, Hicks and, in an apparently very different context, Kalecki, and provided the basis for a theory of the trade cycle which prevailed in the economic profession in the early post-World War II years (the best reference is perhaps Phillips 1954).

Suppose that, while desired and actual consumption are still equal, discrepancies are permitted to exist between demand and supply and between actual and desired investment. We therefore need to replace the relevant equilibrium conditions

$$\dot{Y} = C + I \quad (\text{or equivalently,} \quad sY = I)$$

and

$$I = v\dot{Y}$$

by adjustment mechanisms which reflect economic agents' reactions to undesired situations.

The most commonly used such adjustments are those of a tâtonnement type, according to which the relevant variables change at a rate proportional to the differences between their desired and actual values. In terms of our model, we have

$$\dot{Y} = \tau_y[(C + I) - Y] = \tau_y[I - sY], \qquad (9)$$

$$\dot{I} = \tau_y\left[v\dot{Y} - 1\right], \qquad (10)$$

where τ_y and τ_i, are the (positive) speeds of adjustment of income and investment. The Eq. (9) can be interpreted as a (typically Keynesian) situation in which, prices being fixed and potential supply unlimited, producers are constrained only by demand and adjust their production in relation to (positive or negative) excess demand.

The system (9 and 10) can be easily transformed into a single second-order differential equation in *Y*. By choosing the arbitrary unit of measure of time such that $\tau_i = 1$ we have

$$\ddot{Y} + \left[1 + \tau_y \mathrm{S} = \tau_y \mathrm{v}\right]\dot{Y} + \tau_y sY = 0 \qquad (11)$$

System (11) has a unique position of stationary equilibrium at Y = 0. ('Zero' must be taken here to indicate a level of income determined by factors not considered in the present discussion, such as government expenditure.) Its dynamic behaviour depends on the structural coefficients and may induce decline or growth, in either case with or without fluctuations.

Generally speaking, we may say that the accelerator is an explosive factor in so far as, for given s and τ_y the greater v the more likely it is for the system to grow in time. Moreover, the relative size of the accelerator affects the oscillatory behaviour of the system: if the motion is damped, a large v tends to make the system fluctuate; vice versa, if the motion is explosive, a strong acceleration leads to sustained growth without fluctuations. In agreement with intuitive considerations, large speeds of adjustment tend to produce explosive behaviour, whereas the saving ratio acts as a damper. The effect of these factors on oscillations is more complicated and cannot be ascertained in any obvious way.

A very special and unlikely case arises when we have

$$1 + \tau_y(s - v) = 0 \qquad (12)$$

and the time path followed by the system is a pure sinusoid describing a persistent and perfectly regular cycle, neither damped nor explosive. This of course is a watershed situation which would be destroyed by any small perturbation of the model and is therefore not a suitable idealization of economic cycles.

The multiplier–accelerator model constitutes a rough but effective idealization of certain basic mechanisms deemed to determine or influence cycles and growth in a capitalist economy under certain specific circumstances.

Two major extensions of the model, which have made it theoretically more robust (and complicated), should be mentioned in concluding this entry.

First of all, the assumption that the structural coefficients are constant may be dropped and they may instead be treated as functions of the level (or the rate of change) of income, thus making the model nonlinear. Formal investigation of nonlinear multiplier–accelerator models was initiated in the 1950s by Richard Goodwin (1951a, b) and is still a very active area of research. Nonlinear models have two distinct advantages over the linear ones. For one thing, they better correspond to empirical observation of economic facts. Secondly, and most importantly, they can reproduce a far richer (and economically more interesting) diversity of dynamic behaviours. In particular, only they can represent sustained fluctuations of income, i.e. cycles that neither expire nor explode, without requiring very special configurations of parameters for which no economic justification could be found.

A second important extension of the model has been the generalization of some basic results to the multidimensional case. In an economy with an indefinitely large number of sectors the Harrod–Domar Eq. (7) can be rewritten as

$$[I - A]x = B, \dot{x}, \tag{13}$$

where $A \in R^{nxn}$ is the flow input–output matrix, i.e. the generalized propensity to consume; $B \in R^{nxn}$ is the stock input–output matrix, i.e. the generalized accelerator; $x \in R''$ is the vector of production levels; I is of course the identity matrix.

In analogy to the one-dimensional case we can introduce error-adjustment mechanisms for production and investment, obtaining the system of differential equations

$$\ddot{x} + \left\{T_i + T_y[I - \mathrm{A}] - T_i T_y B\right\}\dot{x} + T_i T_y [I - A]x = 0 \tag{14}$$

where T_y and T_i are diagonal matrices whose (positive) elements are the speeds of adjustment of production and investment, respectively, in the various sectors.

The analysis of system (13) is obviously more complex than that of (11), even in the linear case, as now the coefficients are of order n^2. However, it is possible to define multidimensional equivalents of the main explosive and damping factors, and to indicate the conditions for oscillatory behaviour. It is also possible to show that – in perfect analogy to the one-dimensional case – the loss of stability which takes place when the explosive forces (the accelerators) become too strong vis-à-vis the damping forces (saving and lags), leads to cyclical behaviour of the system.

See Also

- ▶ Acceleration Principle
- ▶ Aggregate Demand Theory
- ▶ Growth and Cycles
- ▶ Multiplier Analysis
- ▶ Trade Cycle

Bibliography

Aftalion, A. 1909. La réalité des surproductions générales. *Revue d'économie politique* 23: 81–117, 201–229, 241–259.

Clark, J.M. 1917. Business acceleration and the law of demand: A technical factor in economic cycles. *Journal of Political Economy* 25: 217–235.

Domar, E.D. 1946. Capital expansion, rate of growth and employment. *Econometrica* 14 (2): 137–147.

Goodwin, R. 1951a. The nonlinear accelerator and the persistence of business cycles. *Econometrica* 19 (1): 1–17.

Goodwin, R. 1951b. Econometrics in business cycle analysis. In *Business cycles and national income*, ed. A.H. Hansen. New York: W.W. Norton.

Harrod, R.F. 1936. *The trade cycle*. Oxford: Clarendon Press.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Hicks, J.R. 1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press.

Kahn, R.F. 1931. The relation of home investment to employment. *Economic Journal* 41: 173–198.

Kalecki, M. 1971. *Selected essays on the dynamics of the capitalist economy 1933–1970*. Cambridge: Cambridge University Press.

Phillips, A.W. 1954. Stabilization policy in a closed economy. *Economic Journal* 64: 290–323.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. *The Review of Economics and Statistics* 21 (2): 75–78.

Multisector Growth Models

Mukul Majumdar

Multisector models are essential ingredients for general equilibrium analysis of an economy over time. They have been used extensively in the literature whenever an adequate description of the relevant issues makes it inappropriate to use aggregative models for formal analysis. A study of optimal accumulation of capital goods or optimal depletion of exhaustible resources is a key to developing a theory of economic planning. The specific results can also be viewed from a different perspective. The idea that markets and prices can be used to achieve efficiency in a decentralized manner has been central to economics. The fundamental theorems of 'new' welfare economics identify conditions under which competitive economies attain an efficient or Pareto optimal allocation of resources. It is natural to enquire whether in dynamic models such a connection between optimality and competitive prices can

be established. One possibility is to use the basic static model and treat the same good at different points of time as different commodities. While such an approach is not entirely shorn of merit, a fundamental paper by Malinvaud (1953) suggested that when economic activity does not terminate at a known date, the outcome of a period-by-period competitive process may fail to be optimal. Indeed, the possibility (or otherwise) of designing an *informationally decentralized* resource allocation mechanism that leads to optimal outcomes has been the subject of considerable speculation and over the last thirty years it has become a 'classical' problem in dynamic models with an infinite horizon. In what follows, I shall review some recent results that throw new light on this topic.

Notation

R^m denotes the m-dimensional Euclidean space; if $x = (x_i) \in R^m$ we write $x \geq 0$ (x is *non-negative*) if $x_i \geq 0$ for all i; $x > 0$ (x is *positive*) if $x \geq 0$ and $x \neq 0$; $x \gg 0$ (x is *strictly positive*) if $x_i > 0$ for all i. $R^m_+ = \{x \in R^m : x \geq 0\}$ and $R^m_{++} = \{x \in R^m : x \gg 0\}$ N is the set of all non-negative integers.

Programmes

There are m producible commodities in the economy (the term 'commodity' is interpreted broadly, including machines of different vintage) and a single non-producible factor of production called 'labour'. Labour is used as an input in production but does not enter into consumption. The supply of labour in period t, denoted by L_t, is given by

$$L_t = L_0 \lambda^t, \quad L_0 > 0, \quad \lambda > 0; \quad t \in N \qquad (1)$$

An activity is a triplet $(L, X, Y) \in R_+ \times R^m_+ \times R^m_+$, where L is the quantity of labour input, X the vector of inputs of producible goods and Y the vector of outputs of producible goods. Let $J' \subset R_+ \times R^m_+ \times R^m_+$ be the set of all technologically feasible activities. The following assumptions on J' are made:

(T′.1) J' is a closed convex cone containing (0,0,0).

(T′.2) '$(L, X, Y) \in J'$, $L' \geq L$, $X' \geq X$, $0 \leq Y' \leq Y$' implies '$(L', X', Y') \in J'$'.

(T′.3) There exists $\left(\widehat{L}, \widehat{X}, \widehat{Y}\right) \in J'$ such that $\widehat{Y} \gg \lambda \widehat{X}$.

(T′.4) $(L, X, Y) \in J'$, $L = 0$ and $Y \neq 0$ implies '$Y < X$'.

(T′.5) '$(L, X_1, Y_1) \in J'$, $(L, X_2, Y_2) \in J'$, $L > 0$, $X_1 \neq X_2$ and $0 < w < 1$' implies that 'there exists $Y > wY_1 + (1 - w)Y_2$ such that $(L, wX_1 + (1 - w)X_2, Y) \in J'$'.

One can interpret the assumptions on J' as follows: (T′.1) means that the technology exhibits constant returns to scale, that inaction is possible and that the production process is continuous (limits of feasible input–output combinations are always feasible). (T′.2) formalizes the idea of free disposal: if (L, X, Y) is feasible, any non-negative output vector smaller than Y is feasible from any input vector larger than (L, X). (T′.3) means that the technology is sufficiently productive: given the natural growth rate λ there is some activity that leads to an increase in per capita stocks. (T′.4) stresses the essential role of labour in the production process: if an activity uses no labour at all, then its net production is non-positive. Finally, (T′.5) is a restrictive strict convexity assumption on the technology.

Given an initial stock $Y_0 \in R^m_{++}$ a *production programme* from Y_0 is a pair of sequences $(\mathbf{X}, \mathbf{Y}) \equiv (X_t, Y_t)_{t \in N}$ such that

$$(X_t, Y_{t+1}) \in J', \quad X_t \leq Y_t \quad \text{for all} \quad t \in N \qquad (2)$$

A production programme generates a consumption programme $C = (C)_{t \in N}$ defined by $C = Y_t - X_t$

It is convenient to use per-capita variables. Define:

$$x_t = X_t / L_t, \quad y_t = Y_t / L_t, \quad c_t = C_t / L_t \quad \text{for} \quad t \in N \qquad (3)$$

Using the assumption (T′.1), we can rewrite the feasibility conditions (2) as:

$$\left(1, x_t, \lambda y_{t+1}\right) \in J', \quad x_t \leq y_t \quad \text{for all } t \in N \quad (4)$$

Define the set

$$JR_+^m \times R_+^m \quad \text{as} \quad J = \{(x, y) : (1, x, \lambda y) \in J'\}$$

Then (4) is equivalent to

$$`x\left(x_t, y_{t+1}\right) \in J, \quad x_t \leq y_t \quad \text{for all} \quad t \in N'$$

Also, one gets

$$c_t = y_t - x_t \quad \text{for all } t \in N.$$

For brevity,

$$(\mathbf{x}, \mathbf{y}, \mathbf{c}) = (x_t, y_t, c_t)_{t \in N}$$

is called a programme from the initial stock y_0. A programme is a complete specification of inputs x_t, outputs y_t and consumptions c_t (measured in per capita terms) in all periods. The relevant constraints are indicated in (4). The set of all programmes from y_0 is denoted by $F(y_0)$.

Evaluation of Programmes

In order to evaluate the welfare-implications of alternative programmes, one introduces an appropriate criterion. With respect to any such criterion, the question of existence of a 'best' or a 'maximal' programme ought to be settled first. In infinite horizon models, subtle consistency problems may arise owing to the fact that an evaluation criterion need not be representable by a real valued function which is continuous in the same topology as that in which $F(y_0)$ is compact. Consider, first, the notion of intertemporal efficiency. A programme **(x, y, c)** in $F(y_0)$ is intertemporally efficient if there does *not* exist another programme $(\mathbf{X}', \mathbf{Y}', \mathbf{C}')$ such that $c_t - c_t \geq 0$ for all t and $c_t' - c_t > 0$ for some t. It is easy to see that $F(y_0)$ contains an infinite number of efficient programmes. A basic question in the literature has been the relation between programmes that are efficient and those that meet the criterion of intertemporal profit maximization at discounted prices. Formally, a programme $(\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{c}})$ in $F(y0)$ is said to satisfy the condition of intertemporal profit maximization if there exists a (non-zero) sequence $\overline{\mathbf{p}} = (\overline{p})_{t \in N}$ of price vectors (in R^m) such that for all $t \in N$ one has:

$$\overline{p}_{t+1}\overline{y}_{t+1} - \overline{p}_t\overline{x}_t \geq \overline{p}_{t+1}y - \overline{p}_t x \quad \text{for } (x, y) \in J \quad (5)$$

Two well-known results (due to Malinvaud 1953) clarify the relationship between efficiency and intertemporal profit maximization.

R.1. *If* $(\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{c}})$ in $F(y_0)$ satisfies the condition of intertemporal profit maximization at prices $\overline{\mathbf{p}} = (\overline{p}_t)$ with $\overline{p}_t \gg 0$ and if

$$\lim_{t \to \infty} \overline{p}_t\overline{x}_t = 0 \quad (6)$$

then $(\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{c}})$ is efficient. It should be emphasized that the 'transversality' condition (6) suggests that a profit-maximizing programme (satisfying (5)) may fail to be efficient due to an over-accumulation of capital inputs. This point has been further explored in Majumdar (1974), Mitra (1976) and Majumdar and Mitra (1976).

Before stating the next result we introduce the notion of non-tightness. A pair $(x, y) \in J$, is non-tight if there exists $(u, v) \gg 0$ such that $(x + u, y + v) \in J$ A programme $(\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{c}})$ is non-tight if (x_t, y_{t+1}) is non-tight for all t. The non-tightness condition requires that an increase of all inputs of producible good leads to an increase in all outputs; roughly speaking, the marginal productivities are all strictly positive.

R.2. *If* $(\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{c}}) \in F(y_0)$ is efficient and non-tight, there exists a non-zero sequence $\overline{\mathbf{p}} = (\overline{p}_t)_{t \in N}$ with $\overline{p}_t \geq 0$ for all t such that $(\overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{c}})$ satisfies the condition of intertemporal profit maximization at $\overline{\mathbf{p}} = (\overline{p}_t)$.

The results R1–R2 can be viewed as an extension of pricing theory characterizing productive efficiency in a static model developed by T.C. Koopmans, who also noted that the condition (6) casts doubts on the feasibility of designing an informationally decentralized resource allocation mechanism that will guarantee efficient outcomes since a verification of (6) cannot be made on the basis of a finite number of observations of $\overline{p}_t$ and $\overline{x}_t$ (Koopmans 1958, pp. 111–27). A formal treatment of this problem has not been available until very recently (see Hurwicz and Majumdar 1984).

A difficulty with the notion of efficiency is that there is an embarrassingly rich class of efficient programme, with widely diverging consumption assignments to different periods. A more precise study of the optimal 'trade-offs' in consumption between two periods can be made if one introduces a one period utility or felicity function. Maximization of a discounted sum of one period utilities generated by consumptions $\mathbf{c} = (c_t)$ has been a well-studied evaluation criterion. However, I shall focus on programmes that are optimal according to Weizsacker's 'overtaking' criterion which avoids discounting. Suppose that consumptions in any period generate utility according to a function $u : R^m_+ \to R$ which satisfies the following properties:

(U.1) u is continuous on R^m_+
(U.2) u is strictly increasing on R^m_{++}
(U.3) u is strictly concave on R^m_+

The continuity property (U.1) means that small changes in consumption levels lead to small changes in utility-levels. The condition (U.2) means that all commodities are desirable; finally, the condition (U.3) formalizes the idea of diminishing marginal utility. It should be stressed that (U.2) and (U.3) do restrict the scope of the model. A programme (x*,y*,c*) in $F(y_0)$ is *optimal* if

$$\lim_{T\to\infty} \sup \sum_{t=0}^{T} \left[u(c_t) - u\left(c_t^*\right)\right] \leq 0 \qquad (7)$$

for all programmes **(x, y, c)** in $F(y_0)$. Convexity of $F(y_0)$ and (U.3) imply that there can be at most one optimal programme. The question of existence has been the subject of extended discussion, and, indeed, a well-known method of proof also establishes an important long-run characteristic ('turnpike' property) of a broad class of programmes. A programme $(\overline{\mathbf{x}},\overline{\mathbf{y}},\overline{\mathbf{c}})$ in $F(y_0)$ is *competitive* if there exists a sequence $\overline{p} = (\overline{p}_t)_{t\in N}$ such that:

$$u(\overline{c}_t) - \overline{p}_t\overline{c}_t \geq u(c) - \overline{p}_t c \quad \text{for all } c \in R^m_+; \quad (8)$$

$$\overline{p}_{t+1}\overline{y}_{t+1} - \overline{p}_t\overline{x}_t \geq p_{t+1}y - p_t x \quad \text{for all } (x,y) \in J \quad (9)$$

A programme **(x, y, c)** is stationary if

$$x_i = x_0, \quad y_i = y_0, \quad c_i = c_0$$

for all $t \in N$. Define

$$D \equiv \{(x,y) \in J : y - x \geq 0\}$$

and

$$C \equiv \left\{c \in R^m_+ : c = y - x, (x,y) \in D\right\}.$$

C and D are non-empty, compact, convex sets. Define $u_* = \max\{u(c): c \in C\}$ There is a unique triplet (x^*, y^*, c^*) such that $(x^*, y^*) \in D, c^* \in C$ and $u(c^*) = u_*$.

For simplicity of exposition I assume that $C^* \gg 0$. The following price support property of (x^*, y^*, c^*) is useful:

R.3. There is $p^* \gg 0$ such that

$$u(c^*) - p^*c^* \geq u(c) - p^*c \quad \text{for all } c \in R^m_+ \quad (10)$$

$$p^*(y^* - x^*) \geq p^*(y - x) \quad \text{for all } (x,y) \in J \quad (11)$$

Using (R.3), one can show that the stationary programme $x_t = x^*, y_t = y_t^*, c_t = c^*$ for $t \in N$ in $F(y^*)$ is competitive: the price system $\mathbf{p} = (p_t)$ is the stationary sequence $p_t = p^*$ (satisfying (8) and

M

(9) for $t \in N$). Furthermore, this stationary programme (known as the golden rule programme) is optimal in $F(y^*)$. Not all competitive programmes from any initial $y_0 \gg 0$ are optimal. The link between optimality (7) and competitive conditions ((8) and (9)) is made precise by the following (for a proof, see Brock and Majumdar 1985):

R.4. Let $c^* \gg 0$ and $y_0 \gg 0$ A programme $(\overline{\mathbf{x}},\overline{\mathbf{y}},\overline{\mathbf{c}})$ in $F(y_0)$ is optimal if and only if it is competitive at prices $\overline{\mathbf{p}} = (\overline{p}_t)$ and

$$V_t = (\overline{p} - p^*)(\overline{y}_t - y^*) \leq 0 \text{ for } t \in N. \quad (12)$$

One can show that *any* competitive programme $(\overline{\mathbf{x}},\overline{\mathbf{y}},\overline{\mathbf{c}})$ satisfying (12) has a 'turn-pike' property:

$$\lim_{t\to\infty} \overline{x}_t = x^*, \quad \lim_{t\to\infty} \overline{y}_t = y^*, \quad \lim_{t\to\infty} \overline{c}_t = c^*.$$

Furthermore, from (10),

$$u(\overline{c}_t) - u(0) \geq \overline{p}_t \overline{c}_t \geq \overline{p}_t (c^*/2)$$

for all sufficiently large t. Since $c = (\overline{c}_t)$ is a bounded sequence, $\overline{p}_t$ is also bounded. Earlier characterization of ('overtaking') optimality of competitive programmes was cast in terms of such a boundedness property of $\overline{\mathbf{p}} = (\overline{p}_t)$ or $(\overline{p}_t \overline{x}_t)$. Once again, whether such sequences are bounded cannot be determined by agents in period t if they are allowed to observe only a finite number of prices and quantities. The verification of (12) by agents in period t requires a knowledge of $\overline{p}_t, \overline{y}_t$ and the vectors p^*, y^* (which can be computed if J is known). Finally, we note that in this model:

R.5. There exists a unique optimal programme from $y_0 \gg 0$.

See Also

- General Equilibrium
- Intertemporal Equilibrium and Efficiency
- Turnpike Theory
- Von Neumann Ray

Bibliography

Brock, W.A., and M. Majumdar 1985. On characterizing optimal competitive programs in terms of decentralizable conditions. Working Paper No. 333, Department of Economics, Cornell University.

Hurwicz, L., and M. Majumdar 1984. Some remarks on optimal intertemporal allocation and decentralization of decisions. Working Paper, Department of Economics, Cornell University.

Koopmans, T.C. 1958. *Three essays on the state of economic science*. New York: McGraw-Hill.

Majumdar, M. 1974. Efficient programs in infinite dimensional spaces: A complete characterization. *Journal of Economic Theory* 7: 355–369.

Majumdar, M., and T. Mitra. 1976. A note on the role of the transversality condition in signalling capital overaccumulation. *Journal of Economic Theory* 13: 47–57.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 23–68.

Mitra, T. 1976. On efficient capital accumulation in a multi-sector neoclassical model. *Review of Economic Studies* 43: 423–429.

Multi-sided Platforms

David S. Evans and Richard Schmalensee
Market Platform Dynamics, Boston, USA

Abstract

This essay provides an overview of the basic economics literature on multi-sided platforms, focusing on ways in which they businesses differ from ordinary single-sided businesses. Distinctive aspects of startup, pricing, welfare analysis, and competition are discussed.

Multi-sided platforms reduce transactions costs and thereby facilitate value-creating interactions between two or more different types of economic agents.

Keywords

Platform; Multi-sided; Two-sided; Network; Externality; Network effects; Critical mass

JEL Codes

D40; L10; L19

Multi-sided platforms (or *MSPs*), which we also call *matchmakers* (Evans and Schmalensee 2016), reduce a transaction cost or economic friction that makes it difficult or impossible for agents in different groups to get together for productive interactions. MSPs are most often created and operated by private for-profit firms, and that's the case on which we focus. Apple's iPhone operating system, for instance, is a two-sided platform linking app developers and consumers. Some MSPs have been the products of non-profit entities (Visa and MasterCard, for instance, were effectively non-profit cooperatives for many years) and a few have been created by governments (this describes national currencies). Some authors, particularly early in the development of the economic literature on MSPs, have labeled them "two-sided markets." Since being an MSP is a property of a market participant, not a market, and MSPs sometimes compete with ordinary one-sided firms, this usage can cause confusion. We have accordingly resisted it.

In some cases by eliminating potential frictions, platforms create opportunities for the emergence of new types of economic agents – app developers for smartphones, for instance. MSPs play critical roles in many economically important industries including payments, communications, financial exchanges, advertising-supported media, operating systems, and various Internet-based industries such as online marketplaces and ride-sharing apps. In many cases, greater involvement by agents of at least one type increases the value of the platform to agents of other types. Such *indirect network effects* function something like economies of scale on the demand side, tending to make larger platforms more attractive to potential customers. A multi-sided platform creates value by coordinating the multiple groups of agents and, in particular, ensuring that there are enough agents of each type to make participation worthwhile for all types.

The fundamental insight that there is a broad class of businesses of this sort that have economic features not well explained by standard textbooks was presented by Rochet and Tirole (2003) in a paper that started circulating around 2000. Other foundational papers are Caillaud and Jullien (2003), Armstrong (2006), and Rochet and Tirole (2006). Weyl (2010) generalizes and unifies the models in these papers. (In the context of information goods, Parker and Van Alstyne (2000) introduced a model that is a linear version of the model subsequently developed by Armstrong (2006).)

The main focus of Rochet and Tirole (2003) was on how the prices charged to the two sides of a platform coordinated demand. They showed that the optimal prices – both from the standpoint of profit-maximization and social welfare maximization – could entail pricing below the marginal cost of provision to one side and above the marginal cost of provision to the other side. Evans (2003a) showed that there were numerous industries in which firms acting as matchmakers set some prices below marginal cost and sometimes at zero.

Since its inception, the literature on multi-sided platforms has grown rapidly in economics, antitrust, and strategic management. In addition, in recent years many new MSPs such as Uber have grown explosively by exploiting advances in computation and communications (Evans and Schmalensee 2016, chapter 3). The multi-sided platform literature is now regularly cited by competition authorities and courts. These businesses pose novel problems for competition policy (Evans 2003b; Evans and Schmalensee 2015).

An Instructive Example

OpenTable is a U.S.-based company that serves restaurants and consumers across the U.S. and in other countries (Evans and Schmalensee 2016, esp. chapter 1). It enables consumers to make and restaurants to accept reservations over the Internet. It helped solve a transaction cost problem for consumers and restaurants. In the U.S., consumers used to have to call a restaurant and, assuming they reached someone, ask whether a particular time was available for their party. If the answer was no, they would repeat the process for another restaurant, perhaps many times. Restaurants used to have to devote resources to taking phone calls, many of which did not result in a

reservation, and keeping track of the reservations they did take.

OpenTable has several features that are common among multi-sided platforms. First, it facilitates valuable interactions between two distinct groups of agents: consumers and restaurants. The fact that members of each group value interacting with members of the other group underlies the indirect network externalities involved and provided an opportunity for an entrepreneur to create a profit-making platform by reducing the transactions costs members of both groups had to incur in order to interact.

Second, OpenTable has three sorts of indirect network externalities. There is a *usage externality*: both consumers and restaurants benefit when each uses the system to make a reservation. And there is a *membership externality*: the system is more valuable to consumers the more different restaurants it lets them access, and the system is more valuable to restaurants the more consumers that use it, since that increases the likelihood that there will be a coincidence between consumers looking for a restaurant and tables available at a particular time.

OpenTable also has what we have called a potential *behavioral externality* (Evans and Schmalensee 2016, chapter 9). Like many MSPs, and unlike most single-sided businesses, OpenTable has rules against conduct that would reduce the value of its platform for other users. In particular, diners who fail to show up for four reservations in a 12-month period have their accounts cancelled.

Third, OpenTable faced a *critical mass* or *chicken-and egg* problem when it began, a problem that is often the most difficult one faced by new matchmakers (Evans and Schmalensee 2010). For OpenTable's service to be viable, it needed to have significant numbers of *both* consumers and restaurants using its platform. OpenTable started by leasing table management software to restaurants – a one-sided business. It developed a web-based platform for consumers to make reservations that linked with its table management software and marketed the online reservation service to consumers for free. After some expensive experimentation, it finally obtained critical mass by working hard to sign up the leading restaurants in a single city, using their presence on the system to market to diners in that city, using that customer base to recruit more restaurants, which then got more diners. It then repeated this formula for obtaining a sufficient density of diners and restaurants that wanted to connect with each other in other cities.

Finally, like many matchmakers OpenTable's price structure involves a *money side*, the group that pays more than marginal cost, and a *subsidy side*, the group that pays less than marginal cost. OpenTable offers its service to consumers for free. In fact, the price to consumers is slightly negative: consumers earn modest usage-based rewards. Restaurants, the money side of this platform, must license Open Table's table management software and pay a fee for every patron they seat who has made a reservation through OpenTable. That is, they pay a fixed *access fee* to be on the platform as well as a *usage fee* when they take a reservation. It is not uncommon for platforms to charge fees of both sorts.

When a single-sided firm with market power sets a price below marginal cost, worries about predatory pricing naturally arise. But OpenTable's charges to restaurants more than cover all its costs. Matchmakers can, of course, engage in predatory pricing, but a correct analysis of their pricing must consider prices to all sides, not just one (Evans 2003b).

To see the complexity of competition policy toward matchmakers, suppose that Open Table proposed a merger with a competitor of roughly equal size and that the merged firm would likely increase prices to restaurants. The merger could still increase the welfare of restaurants, of diners, and possibly both. If restaurants used only one platform and the merged firm did not take the radical step of charging consumers to make reservations, consumers would clearly be better off: they would still face a zero price and could access more restaurants on a single platform. Restaurants might be better off too: they would likely have access to more consumers, and that might more than make up for the price increase.

Critical Mass

All new businesses need to attract enough customers to become profitable before their money runs out. New MSPs also face a chicken-and-egg problem. As noted above, the major challenge for most aspiring platforms is to get enough agents on each side to secure the critical mass necessary to ignite indirect network effects and drive growth (Evans and Schmalensee 2010, 2016, chapter 5). Since platforms are basically selling members of each group access to members of the other group or groups, unless there are enough individuals of the right sort in each group, the platform has nothing to sell. Advertising-supported businesses generally address this chicken-and-egg problem by first producing content to attract audiences, then selling access to those audiences to advertisers.

An interesting example of the power of network effects to ignite growth is provided by Diners Club, the first general-purpose payment card (Evans and Schmalensee 2005, 2007). It gave cards to several hundred consumers in wealthy neighborhoods in Manhattan. It then used that fact to recruit 14 restaurants to take its card. Consumers could use the card for free; restaurants paid a per-transaction usage fee. In the ensuing months more restaurants joined to get access to consumers who wanted to use the card to pay, and more consumers joined to pay at more restaurants. Diners Club ignited. By its first anniversary in 1951, Diners Club had 42,000 individuals who carried its card and 330 merchants that that took the card. Five years later it was accepted at 9,000 merchants, with an annual transaction volume of $54 million. This is an example of what Jullien (2011) has called a "divide and conquer" strategy: subsidize agents in the most price-sensitive group, then use their participation to attract agents in the other group. Other sorts of strategies have also worked for some firms (Evans and Schmalensee 2016, chapter 5).

In contrast, the launch of Apple Pay shows how hard it can be for even sophisticated firms to attain critical mass (Evans and Schmalensee 2016, chapter 10). Apple Pay was launched in October 2014, with rhetoric that promised that it would replace plastic cards at physical points of sale. As of this writing, however, Apple Pay use in the U.S. is insignificant. To use Apple Pay, consumers needed an iPhone 6 or iPhone 6 Plus, and merchants needed to have new terminals that could accept contactless payments. While the new iPhones sold well, most consumers didn't have them at first. Those who did didn't find a compelling reason to use Apple Pay rather than a plastic card, which was easy and convenient. Limited demand to use Apple Pay by consumers gave merchants little reason to acquire the new terminals and promote the use of Apple Pay. Since consumers thus couldn't use Apple Pay at many merchants, even early adopters had little incentive to use it.

Pricing

Pricing in two-sided platforms is more complex than in ordinary multi-product businesses. For single-sided firms, demand depends on the prices of its products as well as the prices of complements and substitutes. For multi-sided platforms, the demand by one group of economic agents also depends on the number of (or, more precisely, measures of the expected value of potential matches with) members of each of the other groups that the platform serves. Loosely speaking, the sides are complements in demand. (Ad-supported media typically require a different analysis because advertisers value more users, but users don't necessarily value more advertising.)

Consider a platform with sides A and B. An increase in price to A-type customers will reduce the number of A's on the platform. Since B-type customers value the platform because of their ability to access A's-type customers, the demand by B's will fall, all else equal. The demand by As will then fall more, since the platform is less valuable to them now that it has fewer B's. As noted by Armstrong (2006), the demand on each side of the platform is more elastic, and the profitability of a price increase is lower, when these positive feedback effects are considered than when they are not considered.

We now briefly consider pricing in the two most basic models of two-sided platforms. In the

first of these, due to Rochet and Tirole (2003), a two-sided monopoly platform operates with no membership externalities, only usage externalities, and levies no membership charges, only per-transaction usage charges. The demand for transactions from group i is given by $D_i(P_i)$, for $i = 1,2$, where P_i is the per-transaction charge to members of group i. One can think of the two groups as merchants and consumers and the platform as a payment system that levies only per-transaction fees. The number of transactions that actually occurs is proportional to the product of the groups' demands in this model, so that, as in real payment systems, there is a value to balanced participation. The platform's profit is given by

$$\Pi = \Big[(P_1 - C_1) + (P_2 - C_2)\Big]\Big[D_1(P_1)D_2(P_2)\Big],$$

where C_i is the per-transaction cost of serving a member of group i.

Let E_i be the (positive) elasticity of D_i with respect to P_i. Then Rochet and Tirole (2003) show that the profit-maximizing prices satisfy the following two optimality conditions:

$$\frac{(P_1 + P_2) - (C_1 + C_2)}{(C_1 + C_2)} = \frac{1}{E_1 + E_2}, \text{and } \frac{P_1}{E_1} = \frac{P_2}{E_2}.$$

The first of these resembles the classic Lerner condition for monopoly equilibrium; the total markup over cost is lower the higher is either demand elasticity. The second condition, however, makes clear that this is not an ordinary multi-product firm. Such a firm would generally maximize profit by charging prices that are inversely related to demand elasticities, all else equal. Here, however, that condition is turned on its head: the optimal prices are *directly* proportional to demand elasticities. Intuitively, the reason is that the platform cares about balanced participation of the two groups, while balance has no value to an ordinary multi-product firm.

In the second basic model, due to Armstrong (2006), a two-sided monopoly platform operates with no usage externalities, only membership externalities, and levies no usage charges, only membership charges. One can think of a heterosexual singles bar in which men value the presence of many women and vice versa. The demand of each group for membership depends both on the fee it is charged and on the number of members of the other group. The firm's profit function in this model is given by

$$\Pi = (P_1 - C_1)D_1(P_1, Q_2) + (P_2 - C_2)D_2(P_2, Q_1),$$

where Q_i is the number of members from group i and $Q_i = D_i(P_i, Q_j)$, $i = 1$, 2 , $i \neq j$.

This model is formally related to the classic model of a monopoly selling complements. In the classic example of coffee and cream, lowering the price of coffee increases the demand for cream because some individuals consume coffee and cream together. Here, however, there are two distinct groups. In the singles bar example, lowering the admission charge to women will increase the demand for admission by men as a reaction to the increased number of women in the bar.

Unlike the Rochet-Tirole (2003) model, the Armstrong (2006) model does not yield simple optimality conditions that hold for all demand functions. Armstrong (2006) shows that in the special case where the D_i functions are linear, the profit-maximizing prices satisfy the following conditions:

$$\frac{P_i - \left(C_i - \theta_{ij}\right)}{P_i} = \frac{1}{\varepsilon_i}, i, j = 1, 2, i \neq j.$$

Here ε_i is the (positive) elasticity of D_i with respect to P_i, holding Q_j constant, and θ_{ij} is a positive term that measures the impact of increases in Q_i on demand from group j , i , $j = 1$, 2 , $i \neq j$. As in the case of complements, prices are lower than they would be in the absence of cross-effects.

Schmalensee (2011) shows that in both these models differences in demand functions can lead to highly skewed pricing of the sort that platform businesses like OpenTable often employ. Weyl (2010) explores a general model that has these two models as special cases, and he shows that they have rather different comparative static properties.

While the Rochet-Tirole (2003) and Armstrong (2006) models form the foundation of much of the multi-sided platform literature, later authors have

introduced additional factors in attempts to produce more tailored models of particular platform types. Hagiu (2009), for instance, modifies the Armstrong (2006) model to capture features of platforms like video game consoles, OpenTable, Amazon, or eBay, that connect differentiated sellers with consumers. He finds that the stronger are consumers' preferences for variety, the larger the share of a monopoly platform's profits that is optimally derived from sellers.

Welfare

An accurate analysis of the impact of any platform's decision on consumer welfare must take into account all the interdependent groups the platform serves. Search engines, for example, provide value to three distinct groups of economic agents: (1) websites that are indexed and made available to people through search queries; (2) people making search queries; and (3) advertisers who are seeking to reach the people who are looking at the search-results page from the query. There are usage and membership externalities across all three groups. The search-engine platform has to balance the interests of these three groups to provide value to them and maximize its own profit. Business decisions that affect the welfare of one group of users are likely to affect the other groups through indirect network externalities. This point is particularly important in the antitrust context, where focusing only on the effects on one group is likely to lead to error.

There are two potential reasons why the profit-maximizing decisions by a platform might differ from the decisions that maximize social welfare. The first is the familiar market power failure. A platform with market power will set its overall *price level* higher than is socially desirable. Since most firms have some market power, the market power failure is not unique to MSPs.

The second possible market failure stems from a platform's choice of its *price structure*. In the two basic monopoly models considered just above, Weyl (2010) shows that this distortion arises because a platform considers the impact of its pricing on the marginal users in the groups it serves, while the impact on the average users is what determines the effect on social welfare. This sort of distortion was first pointed out by Spence (1975) in a model of quality choice by a monopoly. It arises, in principle, whenever a firm with any market power has more than one decision variable and faces buyers who are affected differently by the levels of those variables – that is, almost universally. And, unlike the price level distortion, even its direction depends fundamentally on details of the demand structure: Spence (1975) shows that market-determined quality may be either too high or too low under plausible conditions.

Payment card interchange fees are paid by merchant acquirers (and passed on at least in part to merchants) to bank issuers (and passed on at least in part to consumers). They thus primarily affect the system's price structure. As a very large literature that began with Baxter (1983) makes clear, there is no general reason why the profit-maximizing interchange fee would also maximize social welfare. (See Tirole (2011) for an accessible overview of policy issues and Bedre-Defolie and Calvano (2013) for an interesting recent contribution.) However, the socially optimal interchange fee depends on detailed features of cost and demand structures.

Competition

In simple models, indirect network effects can produce demand-side economies of scale that lead to monopoly: increased participation on one side of the platform makes it more attractive to the other side, leading to increased participation there, making participation by the first side more attractive, and so on. But many of the industries in which indirect network effects are important do not have a single monopoly provider and do not seem to be tending toward monopoly. For example, in the U.S., in addition to several payment systems, there are several competing financial exchanges, numerous magazines even in narrow categories such women's fashion, and multiple shopping malls in most metropolitan areas.

Two features missing from simple models help explain this apparent discrepancy. First, competing

platforms typically offer differentiated products. Second, in some settings customers on one or more sides of the business can patronize more than one platform.

As in one-sided firms, there is often variation among a matchmaker's consumers both in their valuation of various product attributes (horizontal differentiation) and in their willingness or ability to pay for quality (vertical differentiation). For one-sided firms, horizontal and vertical differentiation locates the firm near a pool of potential customers and helps determine pricing. For multi-sided platforms, by determining the customers on one side, horizontal and vertical differentiation affect demand on the other side(s). Because of these interdependencies, a platform must usually make differentiation decisions (including product innovation decisions) jointly for all of the sides it serves. Moreover, the selection of customers on one side is one possible way to differentiate the platform horizontally or vertically.

Product differentiation is a key reason why many industries with multi-sided platforms have multiple competitors. The online portion of the job placement industry, which consists of job boards that help match job searchers with employers through online postings and search, is a highly fragmented industry of two-sided platforms. In the U.S. there are two large job boards that cover many different job categories. But there are also hundreds of other job boards that specialize in different job segments such as professionals (LinkedIn.com) and media jobs (mediabistro.com). By specializing, these job boards presumably increase matching efficiency.

The competitive dynamics of multi-sided platforms depend in theory and in practice on the number of platforms that individual economic agents on each side use, on differences between the two sides in the number of platforms used, and on the ability of an agent on one side to dictate the choice of platform for the other side. Rochet and Tirole (2003) observed that one of the key competitive aspects of multi-sided platforms is the extent to which economic agents engage in what they called *single-homing* or *multi-homing*. An economic agent single-homes if she uses only one platform in a particular industry and multi-homes if she uses several. In the case of payments, consumers and merchants both generally use several payment platforms and therefore multi-home.

Armstrong (2006) showed the importance of multi-homing for competition. Suppose platforms in some market create value by having agents of Type A and Type B as members. If Type A agents only join one platform, then Type B agents can only gain access to Type A agents by joining that same platform. That makes the Type A side of a platform what Armstrong called a *competitive bottleneck*. When there is single-homing on one side and multi-homing on the other side in his model, Armstrong shows that platforms will compete more aggressively for the single-homing customers, who will therefore pay low prices. With these customers on board the platform will then earn its profits from the customers who multi-home on the other side. It is not clear how robust this finding is and how it interacts with other aspects of platform competition. Operating system providers, for example, typically charge users, who single home, and subsidize developers, who multi-home.

Sometimes one set of multi-homing agents can dictate the choice of platform to agents on the other side of the market. Even though most U.S. consumers use multiple payment systems and most merchants accept all of the payment alternatives, one can argue that in practice the consumer dictates which payment system is used. The consumer generally offers one particular payment alternative at checkout. The merchant then has to decide whether to reject that alternative method, with the risk of losing a sale. If the consumer effectively dictates, then, by the logic of competitive bottlenecks, payment platforms have an incentive to compete more aggressively for consumers. Bedre-Defolie and Calvano (2013) show that under this assumption, payment card systems have an incentive to subsidize card users at the expense of merchants.

Multi-sided platforms often face complex competitive environments that involve asymmetric competition (Eisenman et al. 2011). Several common examples include:

- A multi-sided platform competes with single-sided firms on one or more sides. Shopping

malls compete with stand-alone single-sided merchants.
- A multi-sided platform competes on the same sides as a rival but serves an additional side as well. Microsoft Windows competed for users, developers, and computer makers while Apple's MacOS, which wasn't licensed to computer makers, competed only for users and developers.
- Two multi-sided platforms that compete on some but not all sides. This is common for ad-supported media. Facebook operates a social network to attract users to its platform (a two-sided communication network) while Google Search operates a search engine that attracts users looking at search results (from connecting users and websites). Both then connect advertisers to users.

These asymmetries can make both a platform's analysis of its possible decisions and antitrust analysis of platform behavior quite complex. One general lesson for antitrust is that the use of antitrust analytic tools developed for single-sided markets can lead to significant error when applied to MSPs, while multi-sided generalizations of those single-sided tools involve more complexity and information requirements (Evans and Schmalensee 2015).

Conclusions

Just as Molière's bourgeois gentilhomme was surprised and delighted to learn that he had been speaking prose all his life, we and other economists were surprised and delighted to learn from Rochet and Tirole (2003) that the MSPs that we had studied were examples of a large class of businesses that differed in fundamental ways from ordinary, one-sided firms. Managers of MSPs do need to worry about cost and product quality like managers of ordinary businesses, for example, but they also need to worry about balancing participation on all sides of their platforms. This often requires selling below cost to one or more sides or even giving some services away for free, which ordinary firms would never do. Similarly, MSP startups need to attract customers, like all new businesses, but they generally also face a difficult chicken-and-egg problem: they need to attract the right balance of participants. MSPs with market power, like ordinary firms with market power, will generally charge prices above competitive levels. But they may also reduce social welfare by their choice of price structure. And a proper analysis of the welfare effects of any MSP action must consider effects on all sides of the platform.

Because of these and other differences between MSPs and single-sided firms, the antitrust economics of MSPs is more complex than that of ordinary firms. To ignore those differences in the interest of tractability is to risk welfare-reducing policies. Similarly, entrepreneurs, managers, and investors involved with new or established MSPs need to understand their unique features. To ignore them is to invite financial ruin.

See Also

- ▶ Credit Card Industry
- ▶ Online Platforms, Economics Of
- ▶ Network Goods (Theory)
- ▶ Two-Sided Markets

M

Bibliography

Armstrong, M. 2006. Competition in two-sided markets. *RAND Journal of Economics* 37: 668–691.

Baxter, W. 1983. Bank interchange of transactional paper: Legal and economic perspectives. *Journal of Law and Economics* 26: 541–588.

Bedre-Defolie, Ö., and E. Calvano. 2013. Pricing payment cards. *American Economic Journal: Microeconomics* 5: 206–231.

Caillaud, B., and B. Jullien. 2003. Chicken and egg: Competition among intermediation service providers. *RAND Journal of Economics* 34: 309–328.

Eisenmann, T.R., G. Parker, and M. Van Alstyne. 2011. Platform envelopment. *Strategic Management Journal* 32: 1270–1285.

Evans, D.S. 2003a. Some empirical aspects of multi-sided platforms. *Review of Network Economics* 2: 1–19.

Evans, D.S. 2003b. The antitrust economics of multi-sided platforms markets. *Yale Journal on Regulation* 20: 327–379.

Evans, D.S., and R. Schmalensee. 2005. *Paying with plastic: The digital revolution in buying and borrowing*. 2nd ed. Cambridge, MA: MIT Press.

Evans, D.S., and R. Schmalensee. 2007. *Catalyst code: The strategies behind the world's most dynamic companies*. Cambridge, MA: Harvard Business School Press.
Evans, D.S., and R. Schmalensee. 2010. Failure to launch: Critical mass in platform businesses. *Review of Network Economics* 9: 1–26.
Evans, D.S., and R. Schmalensee. 2015. The antitrust analysis of multi-sided platform businesses. In *Oxford Handbook of International Antitrust Economics*, ed. R.D. Blair and D.D. Sokol, 404–447. Oxford: Oxford University Press.
Evans, D.S., and R. Schmalensee. 2016. *Matchmakers: The new economics of multisided platforms*. Boston: Harvard Business Review Press.
Hagiu, A. 2009. Two-sided platforms: Product variety and pricing structures. *Journal of Economics and Management Strategy* 18: 1011–1043.
Jullien, B. 2011. Competition in multi-sided networks: Divide and conquer. *American Economic Journal: Microeconomics* 3: 1–35.
Parker, G.G. and M.W. Van Alstyne. 2000. Internetwork externalities and free information goods. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, 107–216. New York: Association for Computing Machinery.
Rochet, J.-C., and J. Tirole. 2003. Platform competition in two-sided markets. *Journal of the European Economic Association* 1: 990–1029.
Rochet, J.-C., and J. Tirole. 2006. Two-sided markets: A progress report. *RAND Journal of Economics* 37: 645–667.
Schmalensee, R. 2011. Why is platform pricing generally highly skewed? *Review of Network Economics* 10: 1–11.
Spence, A.M. 1975. Monopoly, quality, and regulation. *Bell Journal of Economics* 6: 417–429.
Tirole, J. 2011. Payment card regulation and the use of economic analysis in antitrust. *Competition Policy International* 7: 137–158.
Weyl, E.G. 2010. A price theory of multi-sided platforms. *American Economic Review* 100: 1642–1672.

Multi-valued Treatment Effects

Matias D. Cattaneo

Abstract

The term *multi-valued treatment effects* refers to a collection of population parameters capturing the impact of a treatment variable on an outcome variable when the treatment takes multiple values. For example, in labour training programmes participants receive different hours of training or in anti-poverty programmes households receive different levels of transfers. Multi-valued treatments may be finite or infinite as well as ordinal or cardinal, and naturally extend the idea of binary treatment effects, leading to a large collection of treatment effects of interest in applications. The analysis of multi-valued treatment effects has several distinct features when compared to the analysis of binary treatment effects, including: (i) a comparison or control group is not always clearly defined, (ii) new parameters of interest arise that capture distinct phenomena such as nonlinearities or tipping points, (iii) correct statistical inference requires the joint estimation of all treatment effects (as opposed to the estimation of each treatment effect separately) in general, and (iv) efficiency gains in statistical inference may be obtained by exploiting known restrictions among the multi-valued treatment effects.

Keywords

Causal inference; Generalised propensity score; Identification; Matching estimators; Program evaluation; Semiparametric estimation; Semiparametric efficiency; Treatment effects; Unconfoundedness

JEL Classifications

C14; C21; C31

Treatment Effect Model and Population Parameters

A general statistical treatment effect model with multi-valued treatment assignments is typically described in the context of the classical potential outcomes model. Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) provide recent surveys, with particular emphasis on causal inference in program evaluation. The model

assumes that each unit i in a population has an underlying collection of potential outcome random variables $\{Y_i(t): t \in T\}$, where T denotes the collection of possible treatment assignments. The random variables $Y_i(t)$ are usually called potential outcomes because they represent the random outcome that unit i would have under treatment regime $t \in T$. Each unit is not observed under different treatment regimes simultaneously, which leads to the *fundamental problem of causal inference* (Holland 1986). This idea is formalized in the model by assuming that for each unit i only (Y_i, T_i) is observed, where $Y_i = Y_i(T_i)$ and $T_i \in T$. In words, for each unit i only the potential outcome for treatment level $T_i = t$ is observed while all other (counterfactual) outcomes are missing. Of course, in most applications, which treatment each unit has taken up is not random and hence further assumptions would be needed to identify the treatment effect of interest.

When $T = \{0,1\}$, the model reduces to the classical binary treatment effect model. A finite multi-valued treatment effect model is given by $T = \{0, 1, \ldots, J\}$, for some positive integer J, while $T = [0,1]$ leads to a continuous treatment effect model. (The values in T are ordinal, and may be interpreted as normalisations of the underlying real treatment levels in a given application.) Many applications focus on the classical binary treatment effects model, which has only two groups: *treatment group* ($T_i = 1$) and *control group* ($T_i = 0$). A multi-valued treatment may be collapsed into a binary treatment, which permits the use of classical binary treatment effect (semiparametric) econometric techniques, but this procedure would usually imply an important loss of information in the analysis.

Multi-valued treatment effects are comparisons between some characteristic of the (conditional) distributions of the potential outcomes. Typical examples are mean and quantile comparisons, although in many applications other features of these distributions may be of interest. For example, making the simplifying assumption that the random potential outcomes are equal for all units, the mean of the potential outcome under treatment regime $t \in T$ is given by $\mu(t) = E[Y_i(t)]$. The collection of these means is the so-called *Dose Response Function* in the statistical literature and the *Average Structural Function* in the econometrics literature. Using this estimand, it is possible to construct different multi-valued treatment effects such as pair-wise comparisons $(\mu(t_2) - \mu(t1))$ or differences in pair-wise comparisons, which captures the idea of nonlinear treatment effects. (In the particular case of binary treatment effects, the only possible pair-wise comparison is $\mu(1) - \mu(0)$, which is called the *Average Treatment Effect*.) It is also possible to consider other treatment effects that arise as nonlinear transformations, such as ratios, incremental changes, tipping points or the maximal treatment effect $(\max_{t \in T} \mu(t))$, among many other possibilities. All these multi-valued treatment effects are constructed based on the mean of the potential outcomes, but similar estimands may be considered based on quantiles, dispersion measures or other characteristics of the underlying potential outcome distribution.

Statistical Inference

Identification of (multi-valued) treatment effects is typically achieved by imposing some form of ("local") independence or orthogonality condition together with other model assumptions. A typical identifying assumption is the so-called conditional independence assumption, which assumes that treatment is randomly assigned conditional on a set of observable characteristics. For example, using this assumption, identification is discussed in Imbens (2000) and Lechner (2001) for finite multiple treatments, and in Hirano and Imbens (2004) and Imai and van Dyk (2004) for continuous treatments, while efficient semiparametric estimation of finite multi-valued treatments is studied in Cattaneo (2010). An alternative identifying assumption is an instrumental variables assumption, which assumes the existence of variables that induce exogenous changes in the treatment assignment. For example, using this assumption, Nekipelov (2008) discusses identification and efficient semiparametric estimation of finite multi-valued treatments, while the case of continuous treatments is studied in

Florens et al. (2010). See Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) for comprehensive recent reviews on these and other related results.

See Also

- ▶ Propensity Score
- ▶ Selection Bias and Self-Selection
- ▶ Semiparametric Estimation
- ▶ Treatment Effect

Bibliography

Cattaneo, M.D. 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155: 138–154.

Florens, J.P., J.J. Heckman, C. Meghir, and E.J. Vytlacil. 2010. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica* 76: 1191–1206.

Heckman, J.J., and E.J. Vytlacil. 2007. Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In *Handbook of econometrics, vol. 6B*, ed. J.J. Heckman and E.E. Leamer, 4779–4874. Amsterdam: North-Holland.

Hirano, K., and G. Imbens. 2004. The propensity score with continuous treatments. In *Applied bayesian modeling and causal inference from incomplete data perspectives*, ed. A. Gelman and X.L. Meng. New York: Wiley.

Holland, P.W. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81: 945–970.

Imai, K., and D.A. van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99: 854–866.

Imbens, G. 2000. The role of the propensity score in estimating dose–response functions. *Biometrika* 87: 706–710.

Imbens, G.W., and J.M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47: 5–86.

Lechner, M. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluations of active labor market policies in Europe*, ed. M. Lechner and F. Pfeiffer. Heidelberg: Physica.

Nekipelov, D. 2008. Endogenous multi-valued treatment effect model under monotonicity. *Working paper*, UC-Berkeley.

Multivariate Time Series Models

Christopher A. Sims

The staple of econometrics textbooks, the simultaneous equations model, is a multivariate model; and when the data are time series it becomes a multivariate time series model. John Geweke (1978) laid out the connection of the notation and standard assumptions of simultaneous equations modelling to the corresponding concepts in the theory of vector stochastic processes. Multivariate time series modelling of economic data is none the less a topic distinct from simultaneous equations modelling. We go on to discuss why such a distinction exists, the nature of it, and the prospects for making it less sharp.

Dealing with Over-Parameterization

Both static multivariate models and univariate time series models can easily grow to involve too many unknown parameters. There is usually no obvious limit to how far back in time temporal dependencies might go or on how complex dynamic effects could be, so there is no obvious bound on the number of parameters in a univariate time series model. Every variable in a multivariate model might interact with every other one, meaning there are in the order of n^2 channels of interaction to be parameterized in an n variable model. While this is a finite number, when n is large n^2 can be the same order of magnitude as sample size. When many time series models are modelled jointly, these two sources of parameter proliferation interact multiplicatively.

Classical simultaneous equations methodology takes no explicit account of overparameterization. It presumes there is a model, called the unrestricted reduced form model, for the conditional distribution of the endogenous variables given the exogenous or predetermined variables, and that this model can be estimated.

The focus of the econometric theory is then on how to translate the estimated unrestricted reduced form parameters into efficient estimates of parameters with a more direct economic interpretation, called structural parameters. The number of free parameters is important only in determining whether the model is identified, in the sense that the vector of structural parameters maps into a unique vector of reduced form parameters.

In fact, in a large time series simultaneous equations model it is commonly the case that the unrestricted reduced form model has more parameters than there are data points. Even where this is not true, it is commonly the case that models with nearly as many free structural parameters as there are reduced form coefficients (models which are not strongly over-identified) have far too many parameters.

It is now well understood that classical statistical methods can founder when naively applied to models with too many unknown parameters. Often estimated models are used as certainty equivalents – as if the estimated values of their parameters were known exactly. In this case, applying a false simplifying restriction to a model will reduce expected losses in a decision theory problem if the restriction is not too false and is related to the loss function appropriately. In practice, therefore, econometricians tend to use heavily restricted, simple models, relaxing restrictions when there is enough information in the data to justify doing so.

It is possible to think of inference within approximate, small models whose specification depends on the data as part of a procedure for inference within an infinite dimensional parameter space (see Sims 1972b). Thus in practice it is reasonable to adjust the size of the model in interaction with the data. The problem with the usual implementations of simultaneous equations methodology is not that they make parameterization of the model data dependent, but that they do so while documenting and reporting results according to a theory of inference which ignores the actual specification process.

The various methods grouped under the heading of multivariate time series modelling have in common that they confront the problem of over-parameterization directly. They include prescriptions for how to adjust the model form to obtain a reasonable relation between number of unknown parameters and data points, or in the case of some Bayesian methods a prescription for how to avoid the bad practical consequences of large numbers of parameters without actually reducing the number of them.

By explicitly declaring a strategy for allowing model complexity to depend on the amount and nature of available data, multivariate time series methods open the possibility of separating the application of complexity-controlling model simplifications from the imposition of controversial hypotheses about economic behaviour. Thus time series models may be able to play the role of an unrestricted reduced form where the classical unrestricted reduced form is unusable.

They differ from standard simultaneous equations theory also in that they introduce classes of restrictions on models motivated by hypotheses about the joint dynamic behaviour of the variables. The classical theory focuses attention on analysis of each model equation separately, as a distinct behavioural mechanism. Sometimes knowledge or hypotheses about behaviour do not take this form. Also, restrictions arrived at equation by equation may interact in unexpected ways to imply unreasonable joint behaviour.

Multivariate Time Series modelling Strategies

Static multivariate modelling procedures include principal components, factor analysis, ridge regression, canonical correlation, and multiple-indicator-multiple-cause (MIMC) approaches, among others. Univariate time series procedures include ARIMA modelling (Box–Jenkins), autoregressive modelling and spectral analysis, among others. Multivariate time series modelling multivariate time series models procedures for the most part combine aspects of some well known multivariate modelling strategy with some well known time series modelling strategy, so as to control the dimensionality of the parameter space on the two fronts at once – across variables and across time.

Index Models

Sargent and Sims (1977) introduced a class of models they call index models. If $y(t)$ is a $k \times 1$ vector stochastic process, an index model for it takes the form

$$y(t) + a * z(t) + e(t), \tag{1}$$

where '*' stands for convolution, so that

$$a * z(t) = \sum_{s=-\infty}^{\infty} a(s)z(t-s). \tag{2}$$

The $q \times 1$ vector stochastic process z, the 'index', is taken to have dimension q much less than k, and in most applications interpretation is more natural if $a(s) = 0$ for $s < 0$ (so that only current and past z's influence current y – the z-to-y relation is 'causal' in the jargon of engineering). The model (1) does not determine the properties of the y process, even once a is specified, unless we restrict the joint behaviour of z and e.

One appealing specification is to require that the elements of the e vector be mutually uncorrelated and that the z process be uncorrelated with the e process. Since this implies in general that there is no way to construct current z from current and past x, even if a and the autocovariance function of e are known, Sargent and Sims call this the 'unobservable index' model. It turns out to have the computationally appealing property that, when translated into the frequency domain, it implies that the spectral density matrix at each frequency has the same structure as the covariance matrix of the data in a factor analysis model. Since estimates of the spectral density matrix at sufficiently separated frequencies are independent, the unobservable index model can be treated as a set of independent factor analysis models. There are complications (e.g., the spectral density matrix generally has complex numbers in off-diagonal entries), but the theory of inference for this model is well worked out by Geweke (1977). In this framework, intertemporal parameterization is controlled by the usual frequency domain technique of smoothing the spectral density, while cross-variable parameterization is controlled by keeping q relatively small – keeping down the number of indexes or dynamic factors.

An alternative way to complete the index model is to assert that

$$z(t) = b * y(t-1), \tag{3}$$

with $b(s) = 0$ for $s < 0$, and that $e(t)$ is uncorrelated with $y(s)$ for all $s < t$. Here, of course, z can be constructed from current and past x, so the model is called an 'observable index' model. If we ignore the special nature of the right-hand-side variables, the model is a special case of the MIMC regression model. Also, when we rewrite it as

$$y(t) = (a * b) * y(t-1) + e(t) \tag{4}$$

and recall that e is uncorrelated with lagged y, the model is recognizable as the autoregressive representation for y, linear in y but parameterized so that the coefficients of lagged y's are quadratic functions of a relatively small number of parameters. The model thus combines MIMC and autoregressive modelling as the unobservable index model combines factor analysis and spectral analysis.

State Space Models

In a flexible framework borrowed from engineering, y is modelled as generated by a stochastic 'state vector' z which evolves according to

$$z(t) = Az(t-1) + v(t). \tag{5}$$

The equation for y is

$$y(t) = Hz(t) + e(t). \tag{6}$$

Equation (5) is the 'state equation' and (6) the 'observation equation' in engineering jargon. It is nearly always assumed that e and v are serially uncorrelated and uncorrelated with each other. Because the z vector can be expanded to include lagged values of itself and/or lagged y's, the possible dynamics are rich and the requirement that e and v be serially uncorrelated is not restrictive.

H and A can be allowed to depend on time without affecting the model's tractability.

When v has a full rank covariance matrix, z is just a stochastic process uncorrelated with e. The model is therefore close to the unobservable index model. However, state space models are ordinarily estimated with different techniques. If A and H are known, the Kalman filter provides a convenient method for at the same time finding the likelihood function of the data and forming estimates of the z series. When A and H are unknown, the Kalman filter becomes part of an iterative procedure for choosing A and H to maximize likelihood. Because this iterative procedure is computationally expensive, state space models tend to keep z of small dimension and give A and H simple forms as functions of a small number of parameters. The unobservable index model in the frequency domain, on the other hand, retains its computational tractability only if it is not heavily restricted – otherwise the independence of inference across frequencies is lost.

State space time series modelling is discussed in more detail in Harvey (1981) and, from an engineering perspective, in Kumar and Varaiya (1986).

Bayesian Vector Autoregression

The problem of inference in regression models with large numbers of right-hand-side variables, which can be approached with MIMC models when there are several equations considered jointly, can also be approached with Bayesian techniques or the nearly equivalent 'ridge regression' techniques. Such methods have in fact been applied to univariate or bivariate time series models by Shiller (1973) and Leamer (1972), in a form specialized to take account of a prior belief that coefficients are smaller on more distantly lagged variables and/or similar on adjacent lags. Litterman (1982) suggested a tractable family of specifications for Bayesian prior beliefs about the coefficients in a multivariate autoregression. His specification makes the prior mean for the model a set of possibly correlated random walks, that is a model in which for each i the best predictor of $y_i(t+s)$ based on data up to time t is $y_i(t)$, for all $s > 0$. However, the mean is less important than the covariance matrix of coefficients. His simplest suggestion is to have all coefficients independent, with variances shrinking as the length of the lag increases. He suggests a number of other possibilities as well. Recently, e.g. in Doan, Litterman and Sims (1984), these methods have been extended to allow random time variation in the coefficients. Litterman has published forecasts using a simple model of this type since 1980; the results have been comparable with the performance of commercial forecasting services. They have been relatively better at longer-term forecasts, and they have been relatively good for real variables and relatively bad for prices. See McNees (1986) for more detailed discussion.

Block Structure

A notion which arises in every approach to practical multivariate time series modelling is that of dividing the list y of series into groups corresponding to 'sectors' and limiting the nature of feedback among some sectors. This idea is not in itself a complete modelling strategy, but it is a method for limiting the dimensionality of the parameter space which applies to all time series modelling strategies.

Most commonly, it is assumed that y consists of two subvectors x and z, such that

$$x(t) = f[z(s), x(s-1), s \le t; e(t)], \qquad (7)$$

with $z(s)$ independent of $e(t)$, all s and t. In a usage which has been standard in econometrics for at least 25 years, this condition is called exogeneity of z in the equation (7). I showed in (Sims 1972a) that in the case where (7) is linear with $e(t)$ entering additively, exogeneity of z in (7) implies a restriction on the representation of y as a vector stochastic process which is testable with little in the way of other maintained hypotheses. Since exogeneity assumptions are an important building block in most dynamic simultaneous equations models, their testability is a valuable tool in checking model adequacy.

Economists usually decide which variables are plausibly treated as exogenous by invoking intuitive notions of causal priority. It is therefore useful to observe that the restriction on the joint stochastic process for x and z implied by exogeneity of z in (7) is a causal ordering on x and z, with z first in the ordering, using Granger's (1969) definition of causality.

Causality is a nebulous concept, and Granger's is not a uniquely appealing way to make it precise. None the less, it is formally similar to a number of other proposed precise definitions of causality (see Sims 1977) and has some intuitive appeal.

The statement 'x does not Granger-cause y' is not the same as 'y is Granger-causally prior to x'. Thomas Doan, in an unpublished paper several years ago, observed that if we use Granger's definition of 'x_i causes x_j' in a given multivariate time series model, this relation ('$x_i c x_j$' for short) is not transitive and therefore does not induce a causal ordering on the variables. The relation can be restricted, however, to become transitive. We can treat the relation 'c' as a set of ordered pairs of indexes, so that inclusion of the pair $\langle i, j \rangle$ implies that $x_i c x_j$. We define a new relation 'C' as the largest subset of 'c' which is transitive, that is, satisfies the condition that $x_i C x_j$ and $x_j C x_k$ imply $x_j C x_k$. Then C is well-defined and 'c' can be read as 'Granger-causes' and 'C' as 'is Granger-causally prior to'.

Doan showed that $x_i C x_j$ implies that the complete vector of series in the model, x, can be partitioned into two pieces, one containing x_i and the other containing x_j, such that no variable in the piece containing x_j Granger-causes any of the variables in the piece containing x_i. Thus when these results are specialized to linear time series models, a model with a Granger-causal ordering is one displaying a block triangular structure in its moving average and autoregressive representations. Putting the matter another way, $x_i C x_j$ is equivalent to the assertion that there is some block of variables containing x_i and not x_j which are autonomous, in the sense that the best forecasts which can be made based on past values of variables in the block is as good as the best which can be made when past values of x_j are available as well. This is stronger than the assertion that x_j are available as well. This is stronger than the assertion that x_i does not Granger-cause x_j, which means that the best forecast of x_i itself based on past values of all variables in the x vector other than x_j is as good as the best forecast when past values of x_j are available as well. The assertion $x_i c x_j$ does not by itself connect to an assertion about exogeneity, while $x_i C x_j$ implies that there is some set of equations in x in which x_j appears and x_i is exogenous.

Whatever one thinks of its independent appeal as a definition of causality, Granger causality forces into the open the implicit notion of causal priority underlying exogeneity assumptions in econometrics. Economists often assume exogeneity as a matter of convenience or dogmatism without subjecting these assumptions to critical examination. For example, it is common for models to 'treat as exogenous' policy variables, both because to do otherwise makes use of the models for policy analysis conceptually difficult and because in the policy maker's choice problem there is indeed a sense in which policy variables are 'causally prior'. But treating policy variables as exogenous for purposes of statistical inference amounts to asserting that they are causally prior in Granger's sense. Once we understand Granger's definition, it is easy to see that it is not the same notion of causal priority as that which makes policy variables causally prior in the policy maker's choice problem. Thus the causal priority of policy variables to policy makers does not justify treating those variables as statistically exogenous.

If (7) contains unknown parameters, and if there is another relation to determine z which contains a different set of unknown parameters, then exogeneity of z in (7) implies that estimation of (7)'s parameters in isolation is as efficient as estimating them jointly with the parameters of the relation determining z. Engle, Hendry and Richard (1983) (henceforth EHR) argue that this implication of exogeneity is in fact its essence, arriving thereby at a new definition of the word.

EHR's approach rests on the notion that an economic model is completely characterized by

the function specifying the joint distribution of the data as a function of the parameters. In this framework, a single equation in a model is of separate interest only if it corresponds to a distinct group of parameters. Two sets of equations involving the same parameters and describing the same joint distribution of the data are equivalent. But economists have ordinarily regarded distinct equations or blocks of equations as corresponding to distinct behavioural mechanisms. If our model contains one block of equations determining economic behaviour of Thailand, and another determining economic behaviour of the US, we can ask whether US GNP is exogenous for the model of Thailand. With the standard approach, exogeneity of US GNP depends only on whether there are mechanisms (equations, in a complete model) by which disturbances to Thai economic behaviour influence determination of US economic behaviour. This is an assertion about the true stochastic structure of the world, not about what parameters of the Thai and US models are unknown. But in the EHR approach, if our model contained, say, a parameter representing the rate of technological change in the electronics industry, which was the same in both countries but unknown, no variable from either country would be exogenous in the other.

[A good Bayesian will be suspicious of the distinction I draw here between real randomness in behaviour and uncertainty about parameter values. In a single-agent decision problem this suspicion would be justified. In considering professional communication about scientific inference, the distinction between objective randomness in behaviour and uncertainty about parameters is useful, though (see Sims 1982).]

Another way to see the drawbacks of the EHR approach is to note that in predictive applications of a model, even when the parameters are all known (or more often, when uncertainty in them is ignored), the block structure induced by exogeneity assumptions is a valuable concept. It allows us to characterize patterns of influence in conditional prediction. But the EHR approach leaves the notion of exogeneity undefined when there are no unknown parameters.

The EHR analysis is worth studying, to understand how the usual presumption that analysis conditioning on an exogenous variable can in fact lead to inefficient estimators. This is a subtle point which they illustrate well. But it will remain useful to restrict models by asserting that disturbances in certain equations do not feed back in to the determination of certain variables. This kind of restriction is an assertion about exogeneity in its original sense, not about EHR exogeneity. It seems to me worthwhile to reserve the original sense of 'exogeneity' and to think of EHR exogeneity as a different, related notion.

Granger causal ordering asserts that disturbances in one block of equations do not feed back into determination of variables in a certain block – ever. Sometimes it may not be reasonable to make such an assertion, yet it may be reasonable to assert that the feedback occurs only with a delay with some known lower bound. Block structure based on such feedback delay imposes restrictions on the time series model as does a Granger causal ordering, unless the feedback delay is exactly one time unit.

Feedback delay of one time unit between an equation block and a block of variables in the equations is the assumption of predeterminedness. Predeterminedness is a common assumption both in textbooks and applied work. For much of simultaneous equations theory predetermined variables can play the same role as exogenous variables. In standard set-ups, where each equation has its own set of parameters, predeterminedness coincides with EHR's notion of weak exogeneity.

Because of the arbitrariness of the time unit in economic data, an assertion that intuition or theory tells us there should be a feedback delay of one time unit, but no more, ought to be inherently suspicious. In a model which imposes many predeterminedness assumptions, it ought to be general practice to test for feedback delay of two or more periods. If feedback delays in the model are hardly ever more than one period, the specification ought to be regarded as implausible, even though, since the one-period delay imposes no restrictions in itself, it cannot be tested.

Comparing Modelling Approaches

Each of the modelling strategies discussed above helps to attack the problem of over-parameterization in multivariate time series models. They differ in their amenability to interpretation of various kinds and in their computational tractability.

In models where one has a great deal of a priori knowledge about the dynamic properties of a low dimensional driving process, the state space approach is attractive. It makes using the priori knowledge easy; and in this situation estimates of the historical path of the state process, which emerge naturally in the state space approach, will be important.

The observable index model has many of the same advantages. Its disadvantage is that it has seen little actual use, so that there will be less advice available when the anomalies which crop up so often in applying nonlinear models arise. Also, to the extent that the observable index is kept in a rather general form instead of being specialized in a more or less ad hoc way to a form involving very few unknown parameters (as state space models usually are), it will be computationally less tractable than state space models.

The unobservable index model in the frequency domain is perhaps the most tractable of the approaches discussed here, so long as one limits oneself to testing the hypothesis that a low dimensional index model fits the data and to separating the spectral density into components due to the indexes and due to disturbance terms. It requires considerable additional computational work, however, to generate forecasts and estimates of the historical values of the unobserved indexes. Because it is naturally handled in the frequency domain, it is easy to interpret when the behavioural mechanisms being considered make separate predictions about high and low frequency or seasonal and non-seasonal frequency behaviour.

Bayesian vector autoregressions are computationally easy in some respects, being less dependent on iterative solution methods than the other approaches. However, they avoid the consequences of over-parameterization without actually reducing the dimension of the matrices involved in computation, so that for large models they may make large demands on computer memory. They also do not apply easily to situations where there is *a priori* reason to believe that much of the observed covariation of the data represents common responses to an underlying index or state of low dimension.

The multivariate approaches described above other than block structuring are in themselves symmetric in variables. Only the Bayesian VAR approach explicitly avoids the practice of treating model specifications arrived at by examining the data as if they were actually given a priori. However, index models and state space models do provide a framework for generation of reasonable probability models of multiple time series without invocation of the distinct, theorybased *a priori* knowledge about each equation in the system required with the standard simultaneous equations approach. They therefore share with the Bayesian VAR approach the promise of providing a basis for separating the purely instrumental parts of model specification, which are actually part of the estimation process, from the imposition of restrictions which are grounded in *a priori* knowledge or hypotheses about behaviour.

Imposition of block structure is ordinarily done by a process asymmetric in the variables, strongly influenced by a priori knowledge. We might look for exogeneity of US variables in the model for Thailand, but even if the data seemed compatible with it we would not be likely to impose exogeneity of Thai variables in a model of the US. The same is true of many of the simplifying assumptions routinely invoked in making state space models tractable. In both cases, though, the restrictions differ from the substantive restrictions commonly invoked in standard simultaneous equations modelling in that the connection of the restrictions to the behaviour of the joint probability model for the multiple time series is relatively transparent. In standard models the interaction of hundreds or thousands of restrictions imposed casually on individual equations tends to lead to unexpected anomalies in system behaviour.

Checking Time Series Models for Accuracy

Because there are different approaches to multivariate time series modelling, one commonly is in the position of comparing two or more models for the same data, neither of which is nested in the other. If two models are both nominally unstructured, one may want to know if they fit equally well and whether they are much different. If instead one model is the kind of mathematically intricate, behaviourally simplified model which emerges from modern stochastic dynamic equilibrium theories, while the other is an unstructured multivariate time series model, it is not likely that the equilibrium theory model fits as well as the other, but one may still wish to compare the models. The equilibrium model may provide interpretive insight into the unstructured model if the two are similar in important respects.

Though it might seem so on the surface, this problem of model comparison is not a version of the usual problem of comparing non-nested models. The literature on non-nested models takes the parameterization of each of the models as a firmly maintained hypothesis. In non-Bayesian time series modelling parameterizations are more or less explicitly data-dependent, with more elaborate dynamics emerging when more data are available. And of course the classical literature on non-nested model comparison can be no help at all in comparing a Bayesian to a non-Bayesian time series model.

In comparing model fits, there is a natural measure: recursively generated forecast errors. By these I mean the sequence of forecast errors when the entire modelling strategy is updated at each date t in the sampling period based on data through time t and forecasts made based on the resulting sequence of updated models. Non-Bayesian strategies, though, make the process of using the data to arrive at a model parameterization time-consuming and somewhat subjective, so that it is not practical to reproduce the whole process at each date in the sample. Furthermore, it would be impossible to assure that the researcher's subjective judgements about model form at early dates in the sample were not being influenced by what he knows about the data later in the sample. Bayesian strategies do provide a completely explicit procedure for updating the model based on new data. Even they, however, usually involve some search over a few 'hyperparameters', and this search is seldom reproduced in constructing recursive forecast errors.

There is, unfortunately, no easier alternative to recommend. It is important to remember that even measures of fit nominally based on recursively computed residuals are unreliable when they condition on a particular finite parameterization which has been arrived at after substantial exploration of the data. In fact, in a simple linear regression model it is well known that the sum of squared recursive residuals (weighted by their conditional variances) is the same as the sum of squared 'within-sample' residuals. In other models as well, going through the sample recusively to generate 'out of sample' errors with a model whose parametric form reflects experimentation with the entire sample provides no reliable information about actual out of sample performance.

Some econometricians, recognizing the dependence of the usual specification choice procedures on the data but not willing to abandon or modify those procedures, argue that one can only use actual out of sample performance as a measure of fit. This is a discouraging prescription, however, since historically it is clear that econometric models are revised every few years, so that the best currently available models never have a very extensive record of out of sample forecast performance. And for non-time-series models, there is often no realistic prospect of new data becoming available which would provide a predictive test of the model before decisions based on the results must be taken.

Comparing Time Series Models

It is not always reasonable to compare time series models by comparing how well they fit.

For example one might have available both an uninterpreted Bayesian vector autoregression (VAR) and a fully interpreted behavioural equilibrium model, both applying to the same time series. The latter is likely to contain few parameters and be difficult to solve, but also to be easier to interpret and draw conclusions from. Even if it does not fit as well as the VAR model, we might for some purposes contemplate using it. We would then like to know whether it differs from the betterfitting VAR in substantively important respects.

If the application for the models is known and specific, there will be certain variables or parameters in the model whose conditional distributions given the data are important to the application. Ideally, one compares the models by checking whether they have different implications for these conditional distributions. Where the applications are diverse or ill defined, model comparison becomes correspondingly more difficult.

For models which are linear in variables, the impulse responses, plots of the conditional mean of all variables given unit disturbances to each equation's error term, provide a useful framework for model comparison. If the model is stationary, the impulse responses are also plots of the coefficients of the model's moving average representation and thus a complete summary of the model's second-order properties. They display typical modes of behaviour for variables in the system; they should look qualitatively like plots of the variables being modelled. And they have the units of the variables in the system, so that there is an intuitively reasonable scale for what constitute large differences in impulse responses across models.

For nonlinear models, there is no set of summary measures as appealing as the impulse responses. Where the nonlinearity is not too strong, one can simulate data from the nonlinear models and compare the impulse responses from linear models fitted to the simulated data. Where the focus is on conditional projections and forecasting of the future based on the most recent data, models can be compared by generating conditional distributions of future data with Monte Carlo methods.

Lines of Convergence

It seems likely that the distinction between standard simultaneous equations methods and multivariate time series modelling methods will erode. Simultaneous equations methods, by taking a sophisticated view of dynamic structure and cross-equation serial correlation, can in principle begin to approach multiple time series modelling. Franz Palm (1983) has combined a variant of the Bayesian VAR framework with equation-by-equation simultaneous equations specification.

Keynesian macroeconomic theory connects handily to simultaneous equations econometric theory. It emphasizes separate analysis of consumption, investment, money demand, etc., followed by derivation of conclusions about dynamic behaviour generated by interaction of these distinct mechanisms. Macroeconomic theories based on models of individual optimization under uncertainty tend not to lead to the same clean distinctions among sectors or mechanisms. Empirical analysis of such theories leads restricted multiple time series models to be compared with unrestricted models. In this enterprise classical simultaneous equations theory offers little help. To the extent that this latter type of macroeconomic model becomes more common, emphasis on multiple time series modelling methodology in econometrics is likely to increase.

On the other hand, as multiple time series models are treated more seriously in economics, people will want to use their results. One way or another this forces users of such models to confront the identification problem and thereby is likely to lead to use of formal methods for addressing this problem. A multiple time series model which treats identification formally will in some respects not look very different from a classical simultaneous equations model.

See Also

- ▶ Causal Inference
- ▶ Endogeneity and Exogeneity
- ▶ Forecasting
- ▶ Prediction
- ▶ Time Series Analysis

Bibliography

Doan, T., R. Litterman, and C.A. Sims. 1984. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* 3(1): 1–100; Reply, 131–44.

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

Geweke, J. 1977. The dynamic factor analysis of economic time series. In *Latent variables in socioeconomic models*, ed. D.J. Aigner and A.S. Goldberger. Amsterdam: North-Holland.

Geweke, J. 1978. Testing the exogeneity specification in the complete dynamic simultaneous equations model. *Journal of Econometrics* 7: 163–185.

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

Harvey, A. 1981. *Time series*. New York: Halsted Press.

Kumar, P.R., and P. Varaiya. 1986. *Stochastic systems: Identification, estimation and adaptive control*. New Jersey: Prentice-Hall.

Leamer, E. 1972. A class of informative priors and distributed lag analysis. *Econometrica* 40: 1059–1081.

Litterman, R. 1982. Specifying vector autoregressions for macroeconomic forecasting. In *Studies in Bayesian economics and statistics,*vol. 3 *In Honor of Bruno de Finetti,* ed. P. Good. Amsterdam: North-Holland.

McNees, S. 1986. Forecasting accuracy of alternative techniques: A comparison of US macroeconomic forecasts. *Journal of Business and Economic Statistics* 4: 4–23.

Palm, F. 1983. Structural econometric modeling and time series analysis: An integrated approach. In *Applied time series analysis of economic data*, ed. A. Zellner, 199–233. Washington, DC: US Census Bureau.

Sargent, T.J., and C. Sims. 1977. Business cycle modeling without pretending to have too much a priori economic theory. In *New methods in business cycle research: Proceedings from a conference*, ed. C.A. Sims, 45–109. Minneapolis: Federal Reserve Bank of Minneapolis.

Shiller, R. 1973. A distributed lag estimator derived from smoothness priors. *Econometrica* 41: 775–788.

Sims, C.A. 1972a. Money, income and causality. *American Economic Review* 62: 540–552.

Sims, C.A. 1972b. Distributed lag estimation when the parameter space is explicitly infinite dimensional. *Annals of Mathematical Statistics* 42: 1622–1636.

Sims, C.A. 1977. Exogeneity and causal ordering in macroeconomic models. In *New methods in business cycleresearch: Proceedings from a conference*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Sims, C.A. 1982. Scientific standards in econometric modeling. In *Current developments in the interface: Economics, Econometrics, Mathematics*, 317–337. Dordrecht/Boston/London: D. Reidel.

Mummery, Albert Frederick (1855–1895)

Michael Bleaney

Co-author with J.A. Hobson of the *Physiology of Industry* (1889), Mummery was also a famous mountaineer who wrote a book on climbing in the Alps and the Caucasus and died in the Himalayas in 1895. According to Hobson's own account (*Confessions of an Economic Heretic,* pp. 29–30), it was Mummery who set him on the path to intellectual heresy; considering that Hobson's later economic writings may in many ways be regarded as a development of the theme established in the *Physiology of Industry,* this is a considerable achievement.

Mummery was a businessman who seems to have become acquainted with Hobson by chance while the latter was teaching in Exeter. He managed to convince Hobson, after considerable argument, that the economy contained a serious tendency to over-saving, and that depressions were the expression of this tendency. Unfortunately we do not know how far this idea had developed in Mummery's mind before he met Hobson, or how much each contributed to the published version of the argument. Since Hobson subsequently became a prolific writer on economic matters, one suspects that the meat of the book was his work. It is not certain that Mummery had received much training in economics, and he may have contributed little more than the germ of the idea.

See Also

- ▶ Hobson, John Atkinson (1858–1940)
- ▶ Underconsumptionism

Selected Works

1889. (With J.A. Hobson). *The physiology of industry.* London: Murray.

References

Hobson, J.A. 1938. *Confessions of an economic heretic*. Brighton: Harvester Press, 1976.

Mun, Thomas (1571–1641)

Walter Eltis

Keywords

Balance of payments; Balance of trade; East India Company; Mercantilism; Mun, T.; Specie-flow mechanism

JEL Classifications

B31

Thomas Mun, the distinguished mercantilist, was born in London in June 1571 and died in July 1641. He was the third son of John Mun, a mercer, whose father, also John Mun, held the office of provost of the moneyers in the Royal Mint and received a grant of arms in 1562.

Thomas Mun became an extremely wealthy merchant, and a Director (Member of the Committee) of the East India Company in 1615. In 1624 he had the opportunity to serve as Deputy Governor which he declined, but he remained a director until he died.

The East India Company was much criticized because its trade involved exports of bullion (in order to purchase spices). In 1621 Mun was author of a pamphlet, *A Discourse of Trade, from England unto the East-Indies,* in which he set out the benefits that England derived from this trade. His argument was that the same spices (and he details the amounts) would otherwise have been imported from Turkey at three times the sterling cost, and that the purchase of spices in the Indies thus produced satisfactory results for British consumers, while merchants also benefited, and so ultimately did the balance of trade. On Mun's figures the East India Company exported £100,000 of silver yearly to import silk and spices which sold in England for £500,000 (out of which customs duties took a substantial fraction). But only £120,000 of these goods were actually consumed in England, and the remaining £380,000 were re-exported with the consequence that England gained back considerably more bullion than the original outflow of £100,000.

In 1622 he was the leading member of a committee of merchants which submitted evidence to a Commission set up by James I to investigate the causes of the fall in the exchange rate and the loss of specie from which Britain was suffering. Mun was principal author of their first memorandum in 1622, and sole author of later memoranda submitted in 1623. He strongly opposed Malynes' view that the fall in the exchange rate was attributable to conspiratorial behaviour by foreign merchants, and argued that the balance of *trade* was the principal determinant of specie flows and the exchange rate. His memoranda resurfaced in 1664, as chapters in his posthumously published magnum opus, *England's Treasure by Forraign Trade, or the Ballance of our Forraign Trade is the Rule of our Treasure,* which Schumpeter has referred to as 'the classic of English mercantilism'. This was published by his son, John Mun, with the imprimatur and personal approval of Charles II's Secretary of State, Sir Henry Bennet.

England's Treasure demolished the previous mercantilist literature which advocated detailed interventionist policies to sustain the English money supply and the exchange rate, such as banning gold exports, currency appreciation, lowering the metallic content of the currency, and encouraging the domestic circulation of foreign coin. Mun reiterated the fundamental balance of payments equation that specie flows must be determined primarily by the excess of exports over imports, and therefore insisted that there could not be a sustained loss of gold and silver while there was a trade surplus, while none of the above expedients could prevent a monetary outflow in the face of a sustained deficit.

His book hammered home the significance of the balance of payments equation, with numerous examples to demonstrate the impotence of detailed interventionist policies to hold or attract

bullion while trade was in deficit. At the same time, he developed examples like those in his earlier *Discourse of Trade,* to show how it was ultimately the domestic consumption of imports and not imports as such that needed to be compared with exports to determine the net balance of trade. Imports by English merchants which were not destined for consumption in England were bound to result in equivalent exports, plus of course merchants' profits and duties for the King.

Mun went on to explain the relationship between the balance of trade and the excess of home production over consumption, and to distinguish carefully between the financial interests and impact on the trade balance of Merchants, the Commonwealth (the whole population) and the King. Merchants were solely concerned with profit. The Commonwealth determined the trade balance via the relationship between the aggregate expenditures and incomes of the whole population, while the Sovereign's interest in trade depended considerably upon customs and excise, 'the King by his Customs and Imposts may get notoriously, even when the Merchant notwithstanding shall lose grievously' (p. 147).

Mun may well have been the first to state the celebrated proposition (which Lord Kaldor made much of in the 1970s) that the current account trade surplus must correspond to the sum of the financial surpluses of the public and private sectors. He set out an example where a King enjoys revenues of £900,000, spends £400,000 and accumulates the resulting budget surplus of £500,000. Then if the trade surplus is merely £200,000, the King will.

> lay up £300,000 more in his Coffers than the whole Kingdom gains from strangers by forraign trade: who sees not then that all the money in such a State, would suddenly be drawn into the Princes treasure, whereby the life of lands and arts must fail and fall to the ruin both of the publick and private wealth? So that a King who desires to lay up much money must endeavour by all good means to maintain and encrease his forraign trade. (pp. 188–9)

Mun believed that the achievement of a trade surplus on which monetary inflows depended would be best achieved where the population moderated consumption, and merchants enjoyed maximum freedom to exploit opportunities for trade. He has been much praised in the secondary literature for his perception that it was the trade balance that determined specie flows. This has been universally judged vastly superior to the previous literature which recommended piecemeal interventionism in financial markets. According to McCulloch's (1847) *Edinburgh Review* article 'Mun's book was received as the gospel of finance and commercial policy; and his principles ruled for above a century the policy of England, and much longer that of the rest of Europe' (p. 450).

Mun's analysis was superseded in 18th-century England because he failed to go a vital stage further and appreciate the potentially self-correcting nature of the balance of payments. This led Hume and his followers to cease to regard the trade balance as a primary policy objective in comparison with the achievement of a growing capital stock, and increasing levels of output and employment, about which Mun was also deeply concerned.

But those who have been satisfied that the trade balance is self-correcting have sometimes failed to appreciate that a continuing deficit is inevitable where consumption (modern writers would say, domestic absorption) exceeds production. They also lost Mun's perception that in a protectionist world, winning trade away from other countries may permit increases in domestic capital and employment with would not otherwise occur.

Selected Works

1621. *A discourse of trade, from England unto the East-Indies.* London.

1664. *England's treasure by forraign trade. Or, the ballance of our forraign trade is the rule of our treasure*. London. Repr.in the Economic History Society Reprints of Economic Classics, Oxford, 1928. These works by Mun are both reprinted in J. R. McCulloch, ed., *Early English tracts on commerce.* Cambridge: Cambridge University Press, 1952, which is itself a reprint of the London *Political Economy Club's* 1856 publication, and page references are to this edition.

Bibliography

Appleby, J.O. 1978. *Economic thought and ideology in seventeenth-century England*. Princeton: Princeton University Press.
Heckscher, E.F. 1955. *Mercantilism*, 2 vols. London: Allen & Unwin.
McCulloch, J.R. 1847. Primitive political economy of England. *Edinburgh Review* 172: 426–452.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Supple, B.E. 1970. *Commercial crisis and change in England 1600–42*. Cambridge: Cambridge University Press.
Wilson, C. 1957. *Profit and power: A study of England and the Dutch wars*. London: Longmans.

Mundell, Robert (Born 1932)

Charles Engel

Abstract

Mundell is best known for his work creating an open-economy version of the IS–LM model. Its special interest lies in the analysis of monetary and fiscal policy. He emphasized the importance of the speed of adjustment in capital markets and the role of fixed versus flexible exchange rates in determining the impact of policy changes and the determination of a desirable monetary–fiscal policy mix. Mundell has also been influential on optimum currency areas, the effect of inflation on portfolio investment, and trade theory.

Keywords

Assignment problem; Balance of payments; Central banking; Expectations; External balance; Factor mobility; Fiscal policy; Fixed exchange rates; Flexible exchange rates; Heckscher–Ohlin–Samuelson model; Interest rates; International capital flows; IS–LM analysis; Monetary approach to the balance of payments; Monetary equilibrium; Monetary policy; Mundell, R.; Open-economy models; Optimum currency areas; Sterilization; Sticky prices

JEL Classifications

B31

Robert Mundell is one of the key figures in the development of thought in international monetary economics. His work on the IS–LM model in open economies, equilibrium in a world of perfect capital mobility, monetary dynamics in open economies, and optimal currency areas constitutes the core of the research for which Mundell is best known. His work continues to this day to be influential in the analysis of policy decisions in open economies, but an equally important legacy of Mundell's is the role his work played in determining the direction of research in open-economy macroeconomics in the 1960s, 1970s and through to the present. Mundell's work had such a great impact in part because it combined theoretical rigor with elegant presentation. Mundell was awarded the Nobel Prize in 1999 for 'his analysis of monetary and fiscal policy under different exchange rate regimes and his analysis of optimum currency areas'.

Mundell was born in Kingston, Ontario, in 1932. His undergraduate education was undertaken at the University of British Columbia and the University of Washington. He engaged in postgraduate studies at the London School of Economics and received his Ph.D. from MIT in 1956. He taught at Stanford University and the Bologna Center of the School of Advanced International Studies of the Johns Hopkins University, and joined the staff of the International Monetary Fund in 1961. He was a Professor of Economics at the University of Chicago from 1966 to 1971. In 1974 he joined the faculty at Columbia University, where he has spent the remainder of his career.

Mundell is perhaps best known for his work creating an open-economy version of the IS–LM model. Mundell's (1960; 1961; 1962; 1963a) model is still the workhorse model of most undergraduate texts in international macroeconomics. Mundell, like Meade, Metzler, and a few others whose work preceded Mundell's, recognized that the analysis of exchange rates and balance of payments flows must proceed in a monetary

general equilibrium framework. Under Mundell's initial formulation, the equilibrium conditions in money markets and goods markets were augmented by an external balance condition. Mundell's concept of external balance was a balance of payments equilibrium, in which the net flow demand for foreign exchange is zero. Demand for foreign exchange comes from importers of goods and from importers of foreign assets. In his initial work, Mundell modelled the demand for foreign assets as a flow that depended on the difference between home and foreign interest rates. As long as there was a positive spread between home and foreign interest rates, capital inflows would persist at a steady rate.

Mundell's special interest was in the analysis of monetary and fiscal policy. He emphasized the importance of the speed of adjustment in capital markets and the role of fixed versus flexible exchange rates in determining the impact of policy changes and the determination of a desirable monetary–fiscal policy mix. His framework was extended and used to consider policy issues by academics and central bankers for many years.

One basic insight of these models concerns the difference in the impact of fiscal and monetary policy under fixed and floating exchange rates. Consider a monetary expansion. Under a floating exchange rate, external balance requires a depreciation of the domestic currency. The monetary expansion lowers interest rates, leading to a capital outflow and a decline in demand for the domestic currency. With sticky nominal goods prices (the hallmark of the Keynesian IS–LM analysis), the depreciation makes imported goods more expensive, so expenditure switches to home goods. This expenditure switching effect would not be present if exchange rates were fixed. Indeed, Mundell (1961) makes the point that in the absence of sterilization (see below) the monetary expansion would be reversed over time. That is, under fixed exchange rates, the monetary expansion leads to a balance of payments deficit. Under a balance of payments deficit, as the central bank's foreign reserves decline, the money supply falls.

In contrast, an expansionary fiscal policy might have greater impact under fixed exchange rates, when capital mobility is high. In the IS–LM framework, an increase in aggregate demand raises interest rates. This should lead to an inflow of capital and an appreciation of the home currency under flexible exchange rates. But the appreciation switches demand away from home goods, thereby dampening the effect of the fiscal expansion. Under fixed exchange rates, the expenditure switching does not occur. Moreover, in the absence of sterilization operations the balance of payments surplus that ensues from the fiscal expansion will lead to a domestic monetary expansion as the central bank acquires foreign reserves.

Note how the analysis of the effects of fiscal expansions depends on the assumption that capital flows respond significantly to changes in the interest rate. If capital flows were not significant, the analysis would be reversed. A fiscal expansion leads to an increase in domestic income. Some of that increased income is spent on imports. There may be increased capital inflows because the interest rate has risen domestically, but if these flows are slight then the decline in the trade balance dominates, so the country's balance of payments deteriorates. Under floating exchange rates, then, there will be a currency depreciation that further boosts aggregate demand. That effect is not present under fixed exchange rates, and indeed there could be a contractionary effect of the balance of payments deficit in the absence of sterilization.

Of special note is Mundell's (1963a) version of his model under the assumption of perfect capital mobility, so that the rates of return on home and foreign nominal bonds are equalized. At one level this paper is a simple extension of his earlier work to consider the extreme case in which capital flows infinitely quickly to equalize rates of return. But at another level the model is fundamentally different. In essence this case turns the external balance condition from a flow equilibrium (analogous to the IS curve) into an asset-market equilibrium condition (analogous to the LM curve.) In this model, for asset markets to be in equilibrium households must be satisfied not only with their holdings of money relative to interest-earning assets (LM) but also with their holdings of domestic bonds relative to foreign bonds. This model laid the foundation for virtually all later

work in the field that understands the market for foreign exchange to be an asset market.

The key distinction analytically is that the flow of assets plays no role per se in determining equilibrium in this formulation. For example, the trade balance plays no direct role in establishing the equilibrium in the foreign exchange market. In contrast to many models of the 1950s in which the exchange rate adjusted to set the trade balance to zero, here the trade balance plays a role only in its contribution to the net demand for domestic output. The balance of payments simply reflects the central bank's net accumulation of foreign assets. As Obstfeld (2001) points out, the balance of payments is no longer a relevant indicator of external balance in this setting. By modelling the external balance condition as an asset-market equilibrium, Mundell opened the door for subsequent models that considered the role of expectations in determining exchange rates and laid the foundation for models of balance of payments crises under fixed exchange rates in which speculative attacks play a key role.

Subsequent developments in the field have replaced Mundell's ad hoc formulations of behaviour with optimizing models, and have explicitly modelled expectations formation. But Mundell's work was a cornerstone of the development of more sophisticated models, and open-economy macroeconomic models are still often evaluated by comparing their implications with those of the models of Mundell.

Dynamics was a key concern of Mundell's. Even within the IS–LM framework, Mundell examined the evolution of output, interest rates, exchange rates and prices. Mundell paid special attention to the dynamic effects of balance of payments 'disequilibrium' under fixed exchange rates. When the net private flow demand for foreign exchange is not zero (that is, the sum of the current account and the private component of the capital account is not zero), then, in Mundell's terms, there is balance of payments disequilibrium. Mundell made explicit the distinction between balance of payments flows that were sterilized – so that the monetary base did not change – and policies that allowed the money supply to change automatically when there was balance of payments disequilibrium. Mundell (1961) especially was a precursor of the literature that became known as the 'monetary approach to the balance of payments'. That literature emphasized the automatic adjustment mechanism when there is no sterilization. Most of that analysis was undertaken in classical-style models in which nominal goods prices were assumed to be flexible. Indeed, Mundell (1967) was a contributor in that tradition. But what Mundell's (1961) piece makes clear is that it is the assumption of non-sterilization that is key to understanding the dynamics of adjustment. Even in a world of sticky nominal prices, automatic adjustment to balance of payments disequilibrium can occur through adjustment in the money supply.

Dynamics were central in Mundell's development of what became known as 'the assignment problem'. The question was whether the central bank should be responsible for external balance and fiscal authorities for internal balance, or vice versa. Mundell's answer was that each policy tool should be assigned to the market in which it has the greater effect, which depends on the speed of adjustment of goods markets relative to capital markets. Mundell modelled policymaking in a realistic world in which policymakers have an imperfect understanding of the state of the economy, and in which macroeconomic adjustment to policy changes is slow. These concerns have all but disappeared from more recent research in macroeconomic policymaking, but Mundell's focus still seems relevant. Moreover, Mundell's work recognizes that policymaking at the national level is not in the hands of a single policymaker, but instead involves the interaction of decisions by central banks and fiscal authorities whose actions and goals may not be perfectly coordinated.

Mundell's (1961) paper on optimum currency areas also is still very influential. This paper determines some conditions under which it is optimal for countries to share a common currency. Mundell's view was that there may be some advantage to sharing currencies in terms of reduced transactions costs. But the adoption of a common currency means, of course, that each country is not free to pursue its own independent monetary policy. That loss may not be so large

when factors of production can flow freely between the countries in a currency area. If there is a downturn in one country, adjustment can occur through factor flows towards the country with the stronger economy. But if factor mobility is weak, then there is a case for each country to have its own independent money. In general, in Mundell's framework the optimum currency area is determined by a trade-off between these considerations about factor mobility and considerations involving the transactions costs of having many separate currencies. Mundell's work in this area spawned a large literature that considered other factors that determine whether a set of countries were good candidates for adoption of a single currency.

Mundell is also known for his short paper (1963b) that develops what became known as the 'Mundell–Tobin effect'. Mundell argued that inflation reduced the demand for real money balances. That led to a portfolio shift that could induce greater investment in real capital.

Mundell (1957) also made a lasting contribution in pure trade theory. This paper examined the effects of factor mobility in the Heckscher–Ohlin–Samuelson model.

Factor mobility could be a substitute for goods trade, just as goods trade could substitute for factor mobility (as in the well-known factor-price equalization theorem.)

The Nobel Prize citation notes that 'Mundell chose his problems with uncommon – almost prophetic – accuracy in terms of predicting the future development of international monetary arrangements and capital markets.' When Mundell wrote much of his influential work in the early 1960s, much of the world was on a fixed-exchange rate system – although his native Canada had a freely floating exchange rate. Moreover, there were still significant barriers to international flows of capital that had been erected in the 1930s and 1940s, even among advanced industrialized countries. Nonetheless, Mundell focused in much of his work on the contrast between the fixed and floating exchange rate systems, with an emphasis on the role of capital mobility. Only in the early 1970s did most of the advanced world move to floating exchange rates, and obstacles to capital flows were gradually eliminated in the decades following Mundell's early writings. His work on optimum currency areas was frequently cited in the economic analysis that preceded the introduction of the euro.

Many of Mundell's contributions are collected in *International Economics* (1968). Excellent brief surveys of Mundell's work can be found in Royal Swedish Academy of Sciences (1999) and Obstfeld (2001). Mundell (2001) provides an interesting history of the development of some of Mundell's work.

See Also

- ▶ International capital flows
- ▶ International financial institutions (IFIs)
- ▶ International monetary institutions

Selected Works

1957. International trade and factor mobility. *American Economic Review* 47: 321–355.

1960. The monetary dynamics of international adjustment under fixed and flexible exchange rates. *Quarterly Journal of Economics* 74: 227–257.

1961. A theory of optimum currency areas. *American Economic Review* 51: 657–665.

1962. The appropriate use of monetary and fiscal policy for internal and external stability. *IMF Staff Papers* 9: 70–79.

1963a. Capital mobility and stabilization policy under fixed and flexible exchange rates. *Canadian Journal of Economics* 29: 475–485.

1963b. Inflation and real interest. *Journal of Political Economy* 71: 280–283.

1967. Barter theory and the monetary mechanism of adjustment. In *Capital movements and economic development*, ed. J. Adler. London: Macmillan.

1968. *International economics*. New York: Macmillan.

2001. On the history of the Mundell–Fleming model. Keynote speech. *IMF Staff Papers* 47- (special issue): 215–227.

Bibliography

Obstfeld, M. 2001. International macroeconomics: beyond the Mundell–Fleming model. *IMF Staff Papers* 47- (special issue): 1–39.

Royal Swedish Academy of Sciences. 1999. *Bank of Sweden Prize in Economic Sciences in Memory of Alfred A. Nobel, 1999*. Online. Available at http://nobelprize.org/economics/laureates/1999/ecoback99.pdf. Accessed August 9, 2005.

Municipal Bonds

James M. Poterba

Abstract

Municipal bonds are issued by state and local governments in many nations. In the United States the interest on these bonds is usually exempt from federal income tax, which provides an incentive for taxable investors to hold municipal bonds even if their before-tax yield falls below that of other taxable bonds. This article describes the various types of municipal bonds, the yield spread between taxable bonds and municipal bonds, and the factors that determine the efficiency of the federal income tax exemption as a means of subsidizing capital outlays by state and local governments.

Keywords

Alternative minimum tax (USA); Bonds; Implicit tax rate; Municipal bonds; Social security in the United States

JEL Classifications

H24; H74

Municipal bonds differ from most other securities because of their special tax status. Interest payments on bonds issued by state and local governments in the United States are exempt from federal income tax. Most states with income taxes also exempt their own interest payments from tax. The federal income tax exemption for municipal bond interest is usually justified on the grounds that it reduces borrowing costs for states and localities, thereby facilitating their investment in public infrastructure.

When the federal income tax was enacted in 1913, there was some question as to the constitutionality of such a federal tax on interest paid by states and localities. In 1988, the Supreme Court affirmed the federal prerogative to tax such interest in the case of *South Carolina v. Baker.* The tax exemption for municipal bond interest should therefore be viewed in the same way as any other tax expenditure, namely as a political decision about the structure of income taxation.

There are three types of municipal bonds: *general obligation bonds,* which are backed by the 'full faith and credit' of the borrowing jurisdiction; *revenue bonds,* which are backed by the stream of income from a particular project such as a highway or publicly operated power plant; and *private purpose bonds*, which are tax- exempt bonds issued by private borrowers with the authorization of a state or local government. Only general obligation or 'GO' bonds have a potential claim on the tax revenues of a state and local government. The interest payments on revenue bonds are dependent on the revenues associated with the project that issued the bonds. Private purpose bonds are typically used to finance private sector projects that are deemed beneficial to the state or local economy or community; in practice these bonds finance a wide range of activities. The market value of outstanding tax-exempt bonds in 2006 was 2.3 trillion dollars according to estimates from the Federal Reserve Board Flow of Funds Accounts. GO bonds account for roughly 40 per cent of outstanding tax-exempt debt.

While municipal bond interest is generally exempt from federal income taxation, the relevant tax rules are complicated in some situations. For example, retirees who receive Social Security benefits must include tax-exempt bond interest in the income concept that is used to determine how much of their Social Security income is included in taxable income. In addition, the interest paid on many private purpose bonds is taxable

under the federal alternative minimum tax (AMT). While the AMT affected only 3.5 million taxpayers in 2006, projections suggest that provided there are no changes in the basic structure of the tax, it will apply to more than 20 million taxpayers by 2010. Bonds that are not exempt from the AMT typically offer investors a higher yield than bonds that pay interest that is completely tax exempt.

In part as a result of changes in the tax law, there have been changes over time in the ownership patterns for municipal bonds. Prior to 1986, commercial banks were the primary holders of short-term municipal bonds while households and insurance companies were the primary holders of long-term municipals. The Tax Reform Act of 1986 sharply limited the incentives for banks to hold tax-exempt bonds, and since then the ownership mix has shifted towards households. According to Flow of Funds data for the third quarter of 2006, households were the direct owners of 37 per cent of outstanding municipal bonds. Mutual funds, which are largely owned by households, accounted for another 33 per cent. Commercial banks hold seven per cent, while property and casualty insurance companies hold 14 per cent.

Investors who hold municipal bonds avoid paying income taxes on their interest income, but they pay an 'implicit tax' when the pre-tax interest rate on municipal bonds is lower than that on an equally risky taxable bond. The yield spread between taxable and municipal bonds is often summarized by the *implicit tax rate.* This is the value of θ for which $(1 - \theta)R_T = R_M$ where R_T is the yield on newly issued Treasury bonds and R_M is the yield on prime grade municipal bonds of comparable maturity. This relationship is only satisfied by newly issued taxable and municipal bonds under the assumption that investors plan to hold these bonds to maturity. Poterba (1986) shows that with forward-looking investors, the implicit tax rate measured from current bond yields reflects not just current marginal tax rates on taxable interest but future marginal tax rates as well. For seasoned bonds, the tax treatment of differences between the purchase price of the bond and the par value complicates the calculation of the implicit tax rate. More generally, when investors sell their bonds before maturity, changes in bond prices may result in taxable capital gains or losses. The definition of the implicit tax rate also assumes that Treasury bonds and prime grade municipals are equally risky, an assumption that some might question.

The implicit tax rate on municipal bonds varies across bond maturities at a given point in time, and it varies over time in part as a result of changes in tax rates and tax rules. During the first week of 2007, the interest rate on 30-year GO bonds with an AAA rating was 4.14 per cent, while the yield on a 30-year Treasury bond was 4.59 per cent. The implicit tax rate based on these values is 9.8 per cent, well below the top statutory marginal tax rate on individual investors, 35 per cent. The yield spread between AAA-rated municipal bonds and AAA-rated corporate bonds is larger, but this comparison raises the challenge of risk adjustment. For the same week, the yield on one-year AAA-rated municipals was 3.53 per cent, while that on one-year Treasury bonds was 4.92 per cent. The implicit tax rate at the one year maturity was therefore 28.3 per cent.

One of the challenges in analysing the municipal bond market is explaining why implicit tax rates are substantially below top statutory rates. Chalmers (1998) discusses various potential explanations and rejects the possibility that differential default risk explains this long-standing pattern. The yield curve puzzle has motivated research on the relative pricing of taxable and tax-exempt bonds. Green (1993) argues for moving beyond yield-to-maturity analysis, such as that underlying the foregoing implicit tax rate computations, and developing a more subtle analysis of the tax-exempt bond market.

The key insight in Green (1993) and several subsequent studies is that fully taxable individual investors are unlikely to regard newly issued tax-exempt bonds and newly issued taxable bonds as competitive investment alternatives. If such investors chose to hold taxable bonds, they should do so by holding bonds that generate income in a way that generates less tax liability than a newly issued bond. The opportunities to earn bond returns that face a lighter tax burden are

greater at longer than at shorter maturities, because divergences between the purchase price of a bond and its par value are potentially greater at long maturities. This role of tax- wise investing appears to receive empirical support in yield curve comparisons at different maturities. It may help to explain why implied tax rates in the municipal bond market are often lower for longer-maturity than for shorter-maturity bonds.

Whether the policy of exempting interest on state and local government bonds from federal taxation is an efficient method of encouraging capital formation by states and localities is a long-standing subject of debate. The answer turns on the difference between the implicit marginal tax rate on municipal bonds, which determines the interest saving of state and local government borrowers, and the weighted-average marginal tax rate of municipal bond investors, with weights equal to the tax-exempt interest receipts of each investor. The latter determines the federal government's revenue cost from exempting interest on state and local government obligations from tax. If the revenue cost exceeds the interest saving, it would cost less for the federal government to provide cash transfers to state and local governments equal to the amount of their current interest saving, while taxing interest on their bonds, than to pursue the current policy of tax exemption. In 2002, the weighted-average marginal tax rate for individual investors who received tax- exempt interest was 30.2 per cent. Feenberg and Poterba (1991) describe the calculation of such marginal tax rates. Since the implicit tax rate on 20-year municipal bonds and Treasuries varied between ten and 20 per cent during calendar 2002, the revenue cost of the exemption for households appears to exceed the interest saving for state and local government borrowers.

The progressivity of the federal income tax schedule is a key determinant of the efficiency of policies that exempt interest from taxation. When the yield spread between taxable and municipal bonds is determined by the marginal tax rate of the *lowest* tax rate investor who holds those bonds, but the revenue cost is determined by the weighted average marginal tax rate of the investors who hold municipal bonds, then the efficiency cost of the tax exemption will be greater when the top marginal tax rates affect many but not all municipal bond investors, and when the top rates are substantially higher than the rates on lower-income households.

When investors have access to taxable and tax-exempt bonds of equal risk, market equilibrium should involve investor clienteles in which investors segment themselves according to their tax rates. High tax rate investors should hold tax-exempt bonds, while low tax rate investors should hold taxable bonds. In practice, this separation does not occur. Poterba and Samwick (2003) show that among households that hold tax-exempt bonds, 55 per cent also hold taxable bonds. In contrast, only 15 per cent of the households that hold taxable bonds also hold tax- exempt bonds. There are risks inherent to holding municipal bonds, such as the risk of tax change, that are difficult to hedge and may incline investors to diversify their portfolios. This may explain why most investors who hold municipal bonds also hold taxable bonds.

There are many innovative products in the municipal bond market, including variable rate municipals, insured municipal bonds, and zero coupon tax-exempt bonds. The bonds issued by several large issuers, particularly large states and revenue authorities, trade in active after-markets, but the markets for many smaller municipal bond issues are not very liquid.

See Also

- ▶ Bonds
- ▶ Fiscal Federalism
- ▶ Local Public Finance
- ▶ Tax Expenditures
- ▶ Taxation of Income

Bibliography

Chalmers, J.M.R. 1998. Default risk cannot explain the muni puzzle: Evidence from municipal bonds that are secured by U.S. Treasury obligations. *Review of Financial Studies* 11: 281–308.

Feenberg, D.R., and J.M. Poterba. 1991. Which households own municipal bonds? Evidence from tax returns. *National Tax Journal* 44: 93–104.

Green, R.C. 1993. A simple model of the taxable and tax-exempt yield curves. *Review of Financial Studies* 6: 233–264.

Poterba, J.M. 1986. Explaining the yield spread between taxable and tax-exempt bonds. In *Studies in state and local public finance*, ed. H. Rosen. Chicago: University of Chicago Press.

Poterba, J.M., and A. Samwick. 2003. Taxation and household portfolio composition: U.S. evidence from the 1980s and 1990s. *Journal of Public Economics* 87: 5–38.

Musgrave, Richard Abel (1910–2007)

Peter Mieszkowski

Abstract

Richard Musgrave is best known for his treatise *The Theory of Public Finance* (1959). His most original and lasting contributions are in taxation theory and public goods theory. His work on tax incidence has been the starting point for all subsequent studies on tax burdens by income classes, and he broke new ground by introducing the concept of equal options as the basis for horizontal equity. His separation of budgetary functions into allocation and distribution branches has acquired increased practical significance as much of the expansion of the public sector has consisted of increased transfer payments.

Keywords

Allocative versus distributive budgetary functions; Charitable giving; Choice under uncertainty; Direct and indirect taxation; Excise taxes; Fiscal stabilization; Horizontal equity; Indirect taxation; Merit goods; Musgrave, R. A.; Nozick, R.; Primary and secondary redistribution; Public finance; Public goods; Rawls, J.; Revealed preference; Risk taking; Social justice; Social rights; Tax incidence; Taxation of capital income; Taxation of corporate profits

JEL Classifications

B31

Born in Koenigstein, Germany, Musgrave was educated at Heidelberg (where he obtained a Diplom Volkswirt in 1933) and at Harvard University (where he obtained his Ph.D. in 1937). After serving at the Federal Reserve System in Washington, he held appointments at a number of leading North American universities and ended his formal teaching career at Harvard, where he was Professor Emeritus. He was an economic adviser to a number of governments, headed foreign tax commissions, and served as editor of the *Quarterly Journal of Economics*.

Richard Musgrave is best known for his outstanding treatise *The Theory of Public Finance*, published in 1959 at a time when social expenditures were growing rapidly throughout the industrial world, and when poverty and social justice had become primary policy concerns. This book, which is comprehensive, has served as a fundamental source for scholars and as a teaching reference. In it Musgrave summarizes and extends his original contributions to expenditure theory and the theory of taxation, provides an extensive review of the classical literature in public finance, and includes a thorough discussion of fiscal and monetary policy developed from a Keynesian perspective. One of the great strengths of the book is Musgrave's broad knowledge of the early European masters of public finance, notably Wicksell and Lindahl. By reviewing the classical writers and relating his theory of the public household to their work, Musgrave built an essential bridge between earlier ideas and the development of modern public goods theory.

Musgrave made significant contributions to virtually all areas of public finance. He wrote on the theory of fiscal federalism and revenue sharing, international aspects of taxation, alternative measures of income tax progressivity, land value taxation, the theory of fiscal sociology, and the effects of tax policy on private capital formation, as well as on various aspects of debt and monetary policy. His most original and lasting contributions can be grouped into two categories: taxation

theory, which includes three major contributions, and public goods theory, in particular his theory of the public household.

One of Musgrave's most distinguished contributions to taxation theory is his joint paper with E.D. Domar on the effects of taxes on risk taking (1944). The authors show that taxes on capital income will not necessarily decrease investment in relatively risky ventures once the loss offset provisions of the tax and its income effects are accounted for. In fact, it is quite likely that risk taking will be encouraged by an interest income tax. This article ranks with the half-dozen most influential articles on taxation written since the mid-1950s, and it represents the first application of the theory of choice under uncertainty to taxation. Its conclusions have proved to be quite robust to more general formulations of the theory of risk taking.

Musgrave's second contribution to taxation theory is his theoretical and empirical work on tax incidence. He has developed most of the general concepts currently used in incidence analysis, and, in one of the first general equilibrium analyses, established the fundamental equivalences between direct and indirect taxes and between general and specific factor taxes. These contributions clarify a much confused issue: whether general excise taxes are shifted forwards to purchasers of taxed commodities or backwards to providers of factor services. They also established the importance of both uses and sources aspects of incidence theory.

Musgrave's work on the allocation of tax burden (1951) to different income groups is a basic contribution to applied analysis and has been the starting point for all subsequent studies on tax burdens by income classes. More recently (1974) he refined this earlier work and covered the distributive aspects of expenditures as well as taxes. In another important study, *The Shifting of the Corporation Tax* (1963), with M. Krzyzaniak, Musgrave developed the first econometric estimates of incidence and concluded that the corporate profits tax is shifted forwards, a finding which gave rise to a large literature.

Musgrave extended and refined the normative theory of equitable taxation and its implications for income taxation and the concept of horizontal equity (1959, ch. 8). Later, in 1976, he broke new ground by introducing the concept of equal options as the basis for horizontal equity. Within this framework, two persons are considered to be in equal positions and should be treated equally if they face the same options. Thus, two persons with the same present value of lifetime earnings would be considered equal. One of the important insights of this concept is that under certain assumptions a consumption-based tax system is more equitable than an income tax system: the first treats equals equally while the second discriminates against persons who save relatively more.

The theory of the public household, Musgrave's unifying perspective on public goods, has provided the basis for many of his insights into that fundamental topic. This theory distinguishes between three branches of government – the allocative branch, which provides for social goods and deals with related questions of efficiency, the distribution branch, which modifies the distribution of income as determined by market forces and inheritance, and the stabilization branch, which is concerned with unemployment and overall economic stability.

He stresses that the failure to distinguish between the three different objectives of budget policy will involve unnecessary conflict and inefficient policy design. For instance, different voters may agree on the objective of fiscal stabilization but may fail to enact a proportional cut in taxes in recession if the proposals to combat recession will increase expenditures or change the distribution of income. Hence, one of the practical principles to emerge from the three-budget classification is that expenditure levels and the distribution of income, or tax shares of individual groups, should be determined independently of stabilization objectives. Similarly, the distinction between allocation and distribution leads to the principle that redistribution should be implemented primarily through a tax-transfer process. This will avoid inefficient increases of public expenditures in the name of progressive objectives.

The distinction between allocation and distribution has acquired increased practical

significance as much of the expansion of the public sector has consisted of increased transfer payments: Social Security and publicly financed medical care. Also, in a wide variety of policy areas, from the regulation of the prices of energy resources to efficient congestion-pricing of urban highways, the conflict between allocation and distribution has led to poor policy design, as he predicted. Compensation systems are needed to offset the redistributive effects of efficient allocation policies.

The value of the distinction between allocation and distribution has been enhanced by the work of Robert Nozick and John Rawls on social justice. Nozick has restated and extended John Locke's doctrine that one is fully entitled to the fruit of one's labour. Rawls developed a very different theory based on a communal claim to the output of high-ability persons. However, the claim structure is voluntarily agreed upon through a social compact, as risk-averse individuals, not knowing their future position, agree behind a veil of ignorance to share their income. This contractual approach to distribution is fully consistent with Musgrave's separation between the allocation and distribution branches.

Musgrave distinguishes between primary and secondary redistribution. Primary redistribution is determined by the social rights that entitle the individual to some share of the social product based on membership in the community, rather than on property ownership or labour supplied. Secondary redistribution is voluntary giving that occurs either through private charities or collective provision. Secondary redistribution is Pareto optimal in that the donor derives more satisfaction from providing the gift to the poor than from additional personal consumption.

The mix between primary and secondary redistribution will vary across societies, according to differences in social values. Also, some social rights, or primary redistribution, may be provided in part in the form of goods and services, such as education, training programmes and medical care. This possibility blurs the separation of allocation and distribution functions.

The primary shortcoming of the distinction between allocation and distribution, however, is not the existence of transfers in kind and the subsidization of certain goods, which Musgrave has classified as merit goods (1957, 1959). As stressed by Samuelson, the fundamental issue is that numerous allocations between social and private goods are Pareto efficient, and the choice of an efficient allocation, a task for the distribution branch, has allocative consequences. In a planning solution, then, allocation and distribution are decided simultaneously, not separately.

Musgrave agrees to the formal correctness of this argument but he argues that this approach implicitly assumes that the planner knows individual preferences, and that the question of distribution is dealt with *de novo*. If, however, the distribution of income is determined primarily by market forces and preferences are not known, a pricing rule or voting rule that induces preference revelation must be designed. The determination of the pricing rule is the allocative function of government. The determination of money income, in conjunction with the pricing rule, is the distributive function.

When considered from a broader perspective the separation of budgetary functions into allocation and distribution branches has been invaluable, both as a normative theory and as a description of the way public agencies operate. Experience shows that it is very important to develop coordination between branches of government. Also, Musgrave's three-branch theory clarifies many positive issues, such as the causes of large foreign trade deficits and the demise of central cities in metropolitan areas, as well as the design of policies to deal with these trends.

The establishment of a framework for the systematic solution of fiscal problems is Musgrave's most significant contribution. His work combined theory, institutional and historical information, a deep understanding of prior work and empirical testing. Like a number of other outstanding economists educated during the turbulent 1930s, he emphasized the practical and concrete applications of academic research in the belief that 'intelligent conduct of government is at the heart of democracy', and until the end of his life was an active commentator on

policy issues. A lovely delineation of his views, along with a contrasting perspective, is found in Buchanan and Musgrave (1999); see also his review of the evolution of ideas on fiscal policy (1987).

See Also

- ▶ Horizontal and Vertical Equity
- ▶ Merit Goods
- ▶ Public Finance

Selected Works

1939. The voluntary exchange theory of public economy. *Quarterly Journal of Economics* 53: 213–237.
1944. (With E.D. Domar). Proportional income taxation and risk taking. *Quarterly Journal of Economics* 58: 388–422.
1948. (With T. Thin). Income tax progression. *Journal of Political Economy* 56: 498–514.
1951. (With J.J. Carrol, L.D. Cook, and L. Frane). Distribution of tax payments by income groups: A case study for 1948. *National Tax Journal* 4: 1–53.
1953. On incidence. *Journal of Political Economy* 61: 306–323.
1957. A multiple theory of budget determination. *Finanz Archiv* NS 17: 333–343.
1959. *The theory of public finance: A study in public economy*. New York: McGraw-Hill.
1963. (With M. Krzyzaniak). *The shifting of the corporation tax*. Baltimore: Johns Hopkins University Press.
1974. (With K.E. Case, and H. Leonard). The distribution of fiscal burdens and benefits. *Public Finance Quarterly* 2: 259–231.
1976. ET, OT, and SBT. *Journal of Public Economics* 6: 3–16.
1987. A brief history of fiscal doctrine. In *Handbook of public economics*, ed. A. Auerbach and M. Feldstein. Amsterdam: North-Holland.
1999. (With J. Buchanan). *Public finance and public choice: Two contrasting visions of the state*. Cambridge, MA: MIT Press.

Music Markets, Economics of

Frederic M. Scherer

Abstract

With the growth of economic prosperity, the demise of feudalism, and the weakening of Western religious institutions, markets for music have been transformed radically. By the 19th century, freelance composition and performance endeavours outweighed the employment of musicians by churches and noble courts. Further changes came from the invention of electrical and then electronic means of recording and disseminating music. The emergence of copyright for musical works strengthened economic incentives for the composition of music.

Keywords

Baumol's cost disease; Copyright; Music markets; Superstars

JEL Classifications

L1; Z11

On 15 January 1787, Wolfgang Amadeus Mozart wrote from Prague to a friend in Vienna that 'here [in Prague] nothing is talked about but *Figaro*; nothing is played, tootled, sung or whistled but [Mozart's *Marriage of*] *Figaro.'* Music was ubiquitous, and Mozart was at the time Prague's favorite composer. More than two centuries later, music is played and listened to incessantly, usually through some electronic medium, in homes, shops, automobiles, trains, and on the streets. But the diversity of composers and forms is much greater. And in the means by which music is created and reaches the ears of its countless appreciators, the market institutions have changed radically.

Early History

The history of musical performance is as old as the history of humanity. A seven- hole Chinese flute

has been carbon-dated to the year 7000 BC. Prehistoric tribes celebrated military events and other special occasions with music from drums, horns, flutes, and a variety of stringed instruments. By the Middle Ages in Europe, the professional performance of music was concentrated in the churches, following traditions inherited from the Judaic temple music of King David, and in the residences of the wealthy, especially the nobility. The Roman Schola Cantorum was founded in the seventh century AD to perform what came to be called Gregorian chant. The first Holy Roman Emperor, Charlemagne (d. 814), imported from Rome a delegation of 12 specialists to propagate the correct use of Gregorian chant in northern Europe. Chapels established in residences of the nobility maintained their own cadre of instrumentalists and singers. Competition between Protestant and Roman Catholic denominations during the 16th century led to innovations in the richness of church music, ranging from the eminently singable hymns of Martin Luther to the polyphonic masses of Giovanni da Palestrina. The musicians employed in noble chapels also provided entertainment at dinners and celebrations, and during the Renaissance period wealthier nobles initiated further specialization, maintaining one group of musicians for chapel and another for secular entertainment. During the second half of the 17th century, a kind of 'cultural arms race' emerged among the hundreds of noble courts in Germany, Bohemia, and Austria. Each court competed for prestige through the quality of the musicians and composers it employed to entertain visitors (see Elias 1969; Baumol and Baumol 1994).

As a golden age of classical music dawned in the 17th century, much of Europe was organized along feudal lines. There was an active market for the hiring of promising musicians, who travelled far and wide in search of the best employment opportunities. But once a musician was retained by a feudal lord, at least throughout much of the European continent, he (seldom she) was often bound to continued servitude at the noble's whim and on the noble's terms. Claudio Monteverdi was able to leave his badly paid, demoralizing position with Duke Vincenzo I of Mantua only after his employer's death in 1612. Johann Sebastian Bach was imprisoned for nearly a month in 1717 when he sought to leave the service of the Duke of Weimar. His contemporary Georg Friedrich Händel was advised by friends to reject an employment offer from the King of Prussia (Scherer 2004, p. 94):

> For they well knew, that if he once engag'd in the King's service, he must remain in it, whether he liked it, or not; that if he continued to please, it would be reason for not parting with him; and that if he happened to displease, his ruin would be the certain consequence.

When he was discharged in an economy move during 1769, Niccolò Jommelli was denied permission to take with him copies of the music he had written for the Duke of Württemberg.

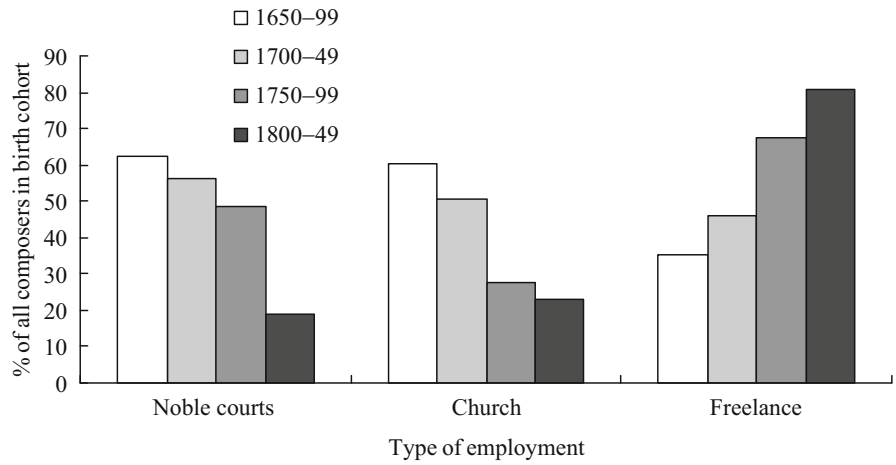
Gradually, however, a new set of opportunities materialized for musicians to earn a living as freelance artists. Opera was the forerunner of this new tradition (see Bianconi and Pestelli 1998). Having pioneered the first modern opera *Orpheo* under ducal auspices at Mantua, Monteverdi migrated to the free city of Venice, where operas were financed by a consortium of wealthy patricians, organized by a hired impresario, and written and performed under contracts individually negotiated with composers, librettists, and soloists. The paradigm spread to other parts of Italy, then to England and parts of Germany, and eventually to other European nations and the United States. Opportunities for the performance of instrumental music at private locales also began to emerge. One predecessor appeared in mercantile London, where King Charles II, embarrassed over his perennial money problems and his inability to pay his court musicians adequately, allowed Henry Purcell and others to perform their music privately in local theaters, taverns, and music halls. In 1697 Thomas Hickford opened a 'Great Dancing Room' in London, perfecting the emerging model for private music halls. In 1735 Vauxhall Gardens, southeast across the Thames from today's Victoria Station, began offering open-air summer concerts at admission prices sufficiently modest to draw Londoners of nearly all economic classes (see McVeigh 1993). These innovations spread

to other locations in London and then to many parts of the European continent. By the third decade of the 19th century, private ballrooms had proliferated in Vienna to the point at which they could accommodate some 50,000 music lovers simultaneously, with entertainment provided, inter alia, by 300 musicians under contract to Johann Strauss, sen., and deployed by Strauss in groups of 25. His son Johann was paid $100,000 to conduct his own and others' compositions at the Boston, Massachusetts, Peace Festival in 1872, performed in a huge wooden shed by an orchestra of 2000 and chorus of 20,000 before audiences of approximately 100,000 persons.

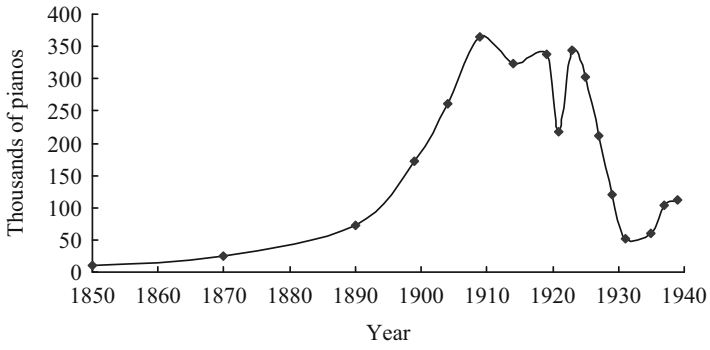
The transition from church and court employment to freelance music composition is depicted in Fig. 1. (Scherer 2004, pp. 69–71). It summarizes by 50-year birth cohort intervals the principal occupational choices of 646 musical composers of enduring fame born between 1650 and 1849. Strong downward trends are evident for court and church employment along with an upward trend for freelance activity. With double-counting allowed to reflect multiple career phases, we see that the fraction employed in noble courts or regularly subsidized by them fell from 62.4 per cent for composers born between 1650 and 1699 to 19.0 per cent for those born in the first half of the 19th century, by which time the Napoleonic wars had undermined much of the feudal system. For church employment the sharpest decline occurs for composers born in the second half of the 18th century. The fraction earning a significant component of their living through freelance composition and performance activities increased from 35.5 per cent for composers born in 1650–99 to 81 per cent for those born in 1800–49.

Music Market Organization

Markets for music are both vertically and horizontally complex. Final demand exists for hearing music performed or for performing it oneself. From that demand are derived a host of other demands: for new musical compositions, for the sheet music through which compositions are disseminated to performers, for training (for example, at conservatories and local schools) in performance, for the concerts and other venues at which music is performed, for the instruments with which it is performed, and for recorded means by which performed music is propagated more widely. The composition, instrument-making, and dissemination stages have for many centuries experienced particularly vigorous innovation. In some subsets, however, such as organ building and violin-making, the technology attained a remarkable degree of perfection as early as the 17th century.



Music Markets, Economics of, Fig. 1 Trends in composers' principal modes of employment, 1650–1849



Music Markets, Economics of, Fig. 2 Trends in US piano production, 1850–1939

Although data permitting a direct statistical test have not been available, the growth of concert-going during the 18th and 19th centuries in tandem with the commercial and industrial revolutions in Europe implies a substantial income elasticity of demand for music consumption. Indirect evidence is presented in Fig. 2 (from Scherer 2004, p. 35), showing trends in the production of pianos in the United States between 1850 and 1939. Values for years other than those on which specific data were available (points) are interpolated. The implied income elasticity of demand is in the range of 2.4–4.3, depending upon what other variables are included in multiple regressions. There is a sharp and lasting break in the series during the mid-1920s, when an economic boom was in full swing. The 1909 and 1923 production peaks were not surpassed over the next 60 years, after which imports began to outweigh domestic production. Two coincident events are responsible for the mid-1920s slump: the introduction of electrical phonographs (with fidelity superior to acoustic phonographs marketed successfully in the 1890s) and the advent of radio broadcasting, including the transmission of classical and popular music. Up to that time, the principal alternative to expensive concert-going (or free summer concerts in urban parks) was the active performance of music within one's home. After the mid-1920s, music could be enjoyed passively at home by listening to radios and phonographs. An era of participatory family musicales began to fade, and a new era dawned.

The marriage of electronics with music wrought further radical changes in markets for music. Through records, radio, television, and still later, the internet, audiences for musical performance were no longer limited to those who could be assembled to hear a specific concert. The whole world was a stage, with four noteworthy consequences. First, for the world as a whole, musical record sales in 1998 (if we count only those sold legally, consistent with applicable copyrights) amounted to more than four billion units. Second, through amplification live performances could be heard by unprecedented numbers of concert-goers. The Woodstock Festival of August 1969 attracted an estimated 300,000–500,000 participants. Third, the expansion of potential audiences enhanced incentives for product differentiation. New musical styles proliferated during the second half of the 20th century at an accelerating pace. Fourth, the prerequisite for success as a vocal performer was no longer a beautiful voice that could carry through the expanse of an opera house. Electronics made popular acclaim attainable for faint voices, and even for performers whose histrionics, dancing ability, costuming, and sex appeal outweighed their vocal talent.

The expansion of markets also intensified a phenomenon already in evidence at the start of the 18th century: superstardom. The received theory (see, for example, Rosen 1981) asserts that the broader the market for talent is, the higher the income differential tends to be between performers with the greatest ability to please and performers of inferior talent. In 1998, for example, the Three Tenors (Luciano Pavarotti, Placedo Domingo, and Jose Carreras) along with their agent received an advance of $18 million for a

single performance accompanying the World Cup football finals in Paris, including broadcast and recording rights. Michael Jackson's '*Thriller*' album, introduced in 1992, achieved an all-time world high of 46 million unit sales, and in 2002, before he plunged into legal and financial difficulties, Jackson's net financial worth was estimated to be in the order of $300 million. But superstardom was not entirely new. In the early 18th century, the leading opera singers made arduous journeys throughout Europe in quest of the most remunerative engagements. The most famous of them all, the castrato Farinelli (Carlo Broschi), is said to have earned £5000 during the 1735–36 opera season in London at a time when an English building craftsman averaged £30 a year.

For live musical performances that are neither amplified nor broadcast, another economic law operates, known as 'Baumol's cost disease' (Baumol and Bowen 1965). Many musical works require a more or less fixed complement of musicians expending a nearly fixed amount of rehearsal and performance time. Thus, labour productivity hardly grows from one century to the next. Meanwhile, most other goods and services experience appreciable rates of productivity growth, permitting those who supply them to earn increasing real incomes over time. For musical performers to stay abreast economically with alternative high productivity growth vocational opportunities, musicians' hourly pay levels must rise apace commensurately, which means that the costs of live musical performances increase relative to the prices of all other goods and services, threatening possibly severe adverse substitution effects. To maintain a thriving supply of live musical performances, subsidization becomes increasingly necessary – not by noble patrons, as in the 18th century, but by governments (preponderantly in Europe and Asia) or affluent concert-goers and private philanthropists (the United States pattern).

For 42 leading US symphony orchestras, all unionized, admissions receipts during the 2002–03 season defrayed on average only 43 per cent of annual budgets. Balancing budgets (which was often not achieved) required voluntary contributions and drawing upon endowments (the latter varying from virtually nothing to $248 million, with a mean of $48 million and median of $19 million). A regression analysis spanning 1980–2002 revealed that those orchestras' budgets were higher, if local population is also taken into account, the greater the concentration of manufacturing, mining, and service corporation headquarters assets was in the relevant metropolitan area. A local corporate headquarters presence subsidized symphony orchestra performance both directly through endowment contributions and through the annual donations of well-paid company officials (Scherer 2005).

Music Publication and Copyright

For at least three centuries the composers and publishers of new music have complained about the unauthorized use, or 'piracy', of their works. The copyright system – having governments confer upon composers and publishers (including record producers) exclusive rights to their productions, which can then be licensed to others upon payment of royalties and/or performance fees – has been the standard means of compromising the maintenance of economic incentives for creative contributions against widespread public dissemination. The first formal copyright law was enacted in England in 1709, but it was interpreted initially not to cover musical works. Extension to musical works came first in 1777 through a lawsuit brought in England by Johann Christian Bach, the son of Johann Sebastian Bach. Musical works were then included under copyright laws passed in the United States, France, various German states, and then, thanks to an initiative led by Johann Nepomuk Hummel and Ludwig van Beethoven culminating in 1837, the German, Austrian, Italian, Czech, and Hungarian territories that previously comprised the Holy Roman Empire.

Prior to the enactment of copyright laws, some protection against unauthorized use was provided by 'privileges' – ad hoc grants of exclusivity conferred upon composers or publishers by royal sovereigns. Securing such grants required access to the relevant sovereign and, in the politically fragmented territories of the old Holy Roman

Empire, the grants were mostly localized and prone to being undermined by competitors producing and selling from another territory. Composers protected their works by contracting with publishers having a reputation for respecting their contracts and keeping manuscripts secret for as long as possible before published versions reached the market. Hand-copying posed a particular threat, for in the early days of the music publishing industry's rapid growth – for example, around 1800 – a copyist could turn out copies by hand at a unit cost lower than the average front-end set-up costs plus variable costs incurred with mechanical printing for production runs of fewer than 25–40 copies (Scherer 2004, p. 162). Like his contemporaries, Mozart attempted to prevent hand copyists from pirating his works by keeping the copyists he hired under constant supervision and dividing work on any given manuscript among multiple copyists. Publishers combatted piracy through secrecy, announcing fixed prices lower than copyists' minimum costs for publications expected to secure a considerable volume (which would now be called 'limit pricing'), keeping composers' honoraria low for works of limited or uncertain appeal, and entering into collusive anti-piracy agreements with fellow publishers.

Giuseppe Verdi and his publisher Giovanni Ricordi were the first to make aggressive use of the copyright laws enacted in German-speaking and Austrian- controlled regions (for example, northern Italy). Previously, local opera houses had purchased or leased manuscripts at cut-rate prices from copyists. With copyright and a network of local enforcement employees, Verdi and Ricordi were able to extract fees from each house, graduating them in a discriminatory fashion to extract more revenue from those serving large, affluent audiences than those located in small provincial towns. They were also particularly energetic in publishing 'reductions' of each separate overture, aria, and chorus, along with bundles covering a full opera, for a diversity of instruments – for example, voice, piano, violin, flute, clarinet, and various ensembles – played by middle-class citizens in their homes. In this way they were able to create a mass market for their works, and as a result Ricordi could pay unprecedentedly large sums for the rights to publish Verdi's works. Verdi became quite rich, accumulating an estate equivalent to nearly £40,000 at the time of his death in 1901 (when English building craftsmen's annual income averaged £100) and beginning semi-retirement at his Busetto villa in the fifth decade of his nearly nine-decade life.

Verdi's extensive written correspondence leaves little doubt that, as his fortune grew, he consciously reduced his work effort along a backward-bending supply curve. Few 18th and 19th century composers achieved as much prosperity as Verdi did; the terminal wealth distribution is highly skew. (Gioachino Rossini became even wealthier and, after reaching the age of 37, spent the remaining four decades of his life in retirement.) It is unlikely that the majority of composers found themselves on the backward-bending portion of a labour supply curve. It is not unreasonable to suppose that the spectacular financial successes achieved by a relatively few composers under the copyright laws inspired many others to try their luck at musical composition. An attempt to test this hypothesis quantitatively (Scherer 2004) was inconclusive, largely because of the difficulty of holding other relevant variables constant. What can be said, however, is that the lack of copyright laws did not prevent classical music from experiencing its golden age of creativity before copyright protection was available in the most musically productive parts of Europe, that is, before the death of Beethoven in 1827 and Schubert in 1828. Despite this limping recommendation, advocates for copyright have been successful in extending greatly both the length of time for which creative individuals and publishers can be protected and, given a continuing stream of new technological challenges, in the range of media over which copyright applies (see Lessig 2004).

See Also

- ▶ Art, Economics of
- ▶ Intellectual Property
- ▶ Superstars, Economics of

Bibliography

Baumol, W., and H. Baumol. 1994. On the economics of musical composition in Mozart's Vienna. *Journal of Cultural Economics* 18: 171–198.

Baumol, W., and W. Bowen. 1965. On the performing arts: The anatomy of their economic problems. *American Economic Review* 55: 495–502.

Bianconi, L., and G. Pestelli (eds.). 1998. *Opera production and its resources*. Chicago: University of Chicago Press.

Elias, N. 1969. *Die höfische Gesellschaft* [The Courtly Society]. Frankfurt: Suhrkamp Verlag.

Lessig, L. 2004. *Free culture*. New York: Penguin.

McVeigh, S. 1993. *Concert life in London from Mozart to Haydn*. Cambridge: Cambridge University Press.

Raynor, H. 1972. *A social history of music from the middle ages to Beethoven*. New York: Schocken.

Rosen, S. 1981. The economics of superstars. *Journal of Political Economy* 71: 845–857.

Scherer, F. 2004. *Quarter notes and bank notes: The economics of music composition in the 18th and 19th centuries*. Princeton: Princeton University Press.

Scherer, F. 2005. Corporate structure and the financial support of U.S. symphony orchestras. Working paper.

Muth, John F. (1930–2005)

F. Robert Jacobs

Abstract

John F. Muth, Professor of Operations Management, is known for his seminal work in rational expectations, aggregate planning and production scheduling. He received his Ph.D. from Carnegie Tech and spent most of his academic career at Indiana University. A colleague for over 20 years, in this article we give insight into his eclectic interests and intellectual motivation.

Keywords

Aggregate planning; Artificial intelligence; Innovation cycles; Linear decision rule; Muth, J. F; Rational expectations

JEL Classification

B31

John F. (Jack) Muth was a brilliant individual, though somewhat awkward socially, and little understood by most people. He was born in Chicago, where his father worked as an accountant at a national accounting firm. Eventually, Jack moved with his parents and two brothers to St. Louis, Missouri. He was very weak as a youngster, suffering from severe asthma and allergies. An avid reader, Jack loved playing the cello and studying mathematics. Jack's cello-playing days continued through the 1980s, and he was a member of the Bloomington symphony orchestra for many years. He studied industrial engineering at Washington University in St. Louis, and continued with graduate work in mathematical economics at Carnegie Tech in Pittsburgh, Pennsylvania. His thesis advisor was Franco Modigliani, with Herb Simon and Merton Miller serving on his committee. All three individuals would later become Nobel laureates in economics.

While a doctoral student, Muth was the first recipient of the Alexander Henderson Award in 1954 (for his work in economics). While finishing his doctorate, he spent the 1957–1958 academic year as visiting lecturer at the University of Chicago, returned to Carnegie Tech as an assistant professor during 1959–1961, spent the 1961–1962 academic year at the Cowles Foundation at Yale University, and finally returned to Carnegie Tech as an associate professor without tenure from 1962 to 1964. It is said that it took him very long to graduate because he did not see the need to take a foreign language examination which would have completed requirements for the Ph.D. degree. Eventually, a colleague whose wife was a French instructor joined the faculty. She tutored Muth in French, and he was finally allowed to graduate. He went on to Michigan State as a professor in 1964, and moved to Indiana University in 1969. He stayed at Indiana University until he retired in 1994.

Throughout his entire academic career John Muth loved to challenge conventional thought. He would explore alternative explanations mathematically, his most famous work being three papers that develop the rational expectations hypothesis (1960, 1961, 1981). Later work by Robert Lucas, the economist, popularized the

idea of rational expectations, and Lucas received the Nobel Prize for his efforts. Esther-Mirjam Sent has written a comprehensive paper that describes Muth's work on rational expectations from a historical perspective (Sent 2002).

Many have asked why Muth did not himself further develop his ideas. If you knew him, this is easy to explain. He knew that there were alternative ways to explain the macroeconomic relationships that were the hot topic of the day. All he wanted to do was show an alternative; in essence, to create an academic debate that challenged conventional wisdom. His colleagues at Carnegie Tech were heavily involved in related research, and he wanted to have some fun and add his thoughts at the same time. Whenever he saw an opening to challenge an idea, he enjoyed developing his elegant ideas, often running concise computer simulations to accompany his mathematical models, subsequently writing up his results. He started doing this very early in his academic career.

A true intellectual, Muth had little interest in promoting his ideas through workshops, presentations or other academic portals. He felt his papers would be interpreted and stand the test of time. Being a good friend, I remember on many occasions Jack talking about invitations to speak at international conferences and at schools. These invitations were usually declined, I am sure not because he was uninterested but rather because he felt these activities would take a significant amount of time, he would probably have trouble with his allergies, and he was more interested in working on his current ideas. He always had something that he was actively working on and would talk about these ideas often over a few beers in the late afternoon at Nick's in Bloomington, Indiana, with his friends.

The late 1960s and early 1970s were spent developing industrial scheduling theory in the field of operations management. He wrote about the importance of the 'aggregate planning' problem and established it in the literature in 1960 with his colleagues at Carnegie Tech (Holt et al. 1960). It was Muth who established the proof of the linear decision rule in aggregate planning. This effort developed into a series of books published with Gerry Thompson and Gene Groff (Muth and Thompson 1963; Groff and Muth 1969, 1972).

He spent the late 1970s through the early 1980s studying artificial intelligence. His main interests were in inference engines and inductive and deductive logic. To my knowledge he published only one paper on the topic (Jacobs et al. 1991). I often heard him refer to his work on artificial intelligence as his 'ten-year sink hole'.

Later in the 1980s he began studying innovation cycles. He would often muse on the fact that many of the most innovative ideas were developed by individuals working at home, and how corporations that spent gigantic sums to develop new ideas so often produced only minor incremental innovations. He wrote simple simulation programs that simulated a random progress function, and matched these results with what was documented in the literature, often musing on the fit. He published an important paper in this area in 1989.

During the late 1980s until his retirement in 1994, Muth spend his time teaching undergraduate courses in process design and scheduling. As one might imagine, he was an awkward teacher and often had difficulty coming down to the level of doctoral students, much less undergraduate students. When he realized during this time that he was going to have to teach undergraduate students to see out his career, it was interesting to observe how he worked to improve. He worked with the teaching resource group at the university, who videotaped his lectures and helped him develop a better teaching style. His colleges in the department were amazed when he was listed as a recommended instructor in the student newspaper in the early 1990s, an event that gave him great personal satisfaction.

He loved sailing in the Florida Keys and had a 30-foot Auburn sailboat that was docked in Marathon until he moved it to Cudjoe Key around 1989. The boat was well suited for sailing around the Florida Keys having a shallow keel. He retired in 1994 and initially split his time between Bloomington and the Keys. For a time, he worked as a consultant to the business school to develop the integrative cases used in the undergraduate core curriculum, taking Indiana's integrative

core to yet another level, as he had done in economics and in almost any area in which he became involved. From around 2000, he remained permanently in the Keys.

This article has possibly not emphasized enough the impact of Muth's work. He was truly a brilliant intellectual and influenced numerous doctoral students throughout his career at Michigan State and Indiana University. As an aside, he was also an amazing Trivial Pursuits player. You always wanted Muth on your team since he seldom missed a question! Muth also had a private side as an aggressively loyal and caring person to his close friends. He was always willing to spend time to talk through important career decisions, and always willing to comment quickly and brilliantly on a manuscript (although it might cost you a beer).

Finally, a funny story, I can still remember being in the Keys with my wife and children, and visiting Jack when he first bought the Cudjoe Key house. Late one afternoon we were all driving up from Key West with Jack. We stopped at a store to pick up some food for dinner, my wife and I leaving the kids with Jack in the car. When we returned to the car, there we saw Jack teaching our two young daughters how to make 'unusual noises' by putting their hands over their armpits and pumping their arms up and down. We all still laugh when we think about that time and the other wonderful times we enjoyed with that nervous little genius who was such a great friend.

Selected Works

1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55, 299–306.

1960. (With C.C. Holt, F. Modigliani and H.A. Simon.) *Planning production, inventories, and work force*. Englewood Cliffs: Prentice-Hall.

1961. Rational expectations and the theory of price movements. *Econometrica* 29, 315–335.

1963. (With G.L. Thompson.) *Industrial scheduling*. New York: Prentice-Hall.

1969. (With G.K. Groff.) *Operations management: Selected readings*. Homewood: Irwin.

1972. (With G.K. Groff.) *Operations management: Analysis for decisions*. Homewood: Irwin.

1981. Estimation of economic relationships containing latent expectations variables. In *Rational expectations and econometric practice*, ed. R.E. Lucas and T.J. Sargent. Minneapolis: University of Minnesota Press.

1986. Search theory and the manufacturing progress function. *Management Science* 32, 948–962.

1989. A stochastic theory of the generalized Cobb–Douglas production function. In *Cost analysis applications of economics and operations research*, ed. T.R. Gulledge, Jr. and L.A. Litteral. New York: Springer-Verlag.

1991. (With F.R. Jacobs, T. Hancock, and K. Mathieson.) A rule–based system to generate NC programs from CAD exchange files. *Computers & Industrial Engineering* 20, 167–176.

Bibliography

Sent, E.-M. 2002. How (not) to influence people: The contrary tale of John F. Muth. *History of Political Economy* 34: 291–319.

Myopic Decision Rules

Mordecai Kurz

In a dynamic context a decision maker at any instant t has information about his exogenous economic environment both at time t and at later dates. We represent the environment at t by a vector $x(t)$ of exogenous variables, and their future values by $(x(t+1), x(t+2), \ldots, x(t+T))$. The horizon T is determined by such considerations as length of life, technology, resource limitations etc.; it might be infinite. A decision rule at time t is a map ψ_t associating with a vector of variables

z the variable d representing the choice of the decision maker. We write $d = \psi_t(z)$. *Myopic decision rules* refer to those maps of the form $d(t) = \psi_t(x(t))$ in which $d(t)$ depends only upon the values of the exogenous variables at time t, disregarding any information about future conditions of the economic environment. A decision rule is said to be *non-myopic* if it is of the form $d(t) = \psi_t(x(t), x(t=1), \ldots, x(t+T))$.

As an example, consider the consumer who wants to maximize the utility function $U_t(c(t), c(t+1), \ldots, c(t+T))$ subject to a budget constraint defined by a vector $(\omega(t), \omega(t+1), \ldots, \omega(t+T))$ of endowments and $(p(t), p(t+1), \ldots, p(t+T))$ of prices. A consumption function like $c(t) = \phi_t(\omega(t), p(t))$ is a myopic choice function whereas a decision function like $c_t = \phi_t(\omega(t), p(t), \omega(t+1), p(t+1), \ldots, \omega(t+T), p(t+T))$ is non-myopic. It is clear from these definitions that myopic decision rules ignore all intertemporal substitution possibilities which may arise from uneven resource distribution, changing needs or prices over time, whereas non-myopic decision rules call upon the decision maker to consider simultaneously all his limitations over the entire relevant period (t, $t+T$) and make optimal intertemporal substitutions based on his constraints.

Historical Review

Non-myopic behaviour of firms was the standard means by which capital theory was developed in the 19th and early 20th century. The typical model of the firm identified it with an investment programme and assumed that the firm seeks to maximize the present value of its profits by selecting an optimal stream of actions. In formulating the problem, the optimal decision of the firm at any date depends upon endowments, prices and technology at all dates. A far more complex view was taken of the consumer. With incompletely developed utility theory, economic theorists in the late 19th century developed only implicit non-myopic decision rules. Böhm-Bawerk (1891) clearly analysed the intertemporal choices of a non-myopic consumer and his 'grounds' for time preference and – although confusing technology, preferences and equilibrium conditions – clearly attempted to identify the preferences which would lead to non-myopic decisions of a consumer. Non-myopic consumers may be found in the writings of most early capital theorists; however, the most complete early formulation of non-myopic decisions of consumers and firms was provided by Fisher (1930). It is his model that has remained the foundation of most discrete time models of intertemporal allocations.

Non-myopic decision models of economic agents arise almost always in the context of microeconomic analysis. It is noteworthy that the greatest thrust of myopic decision rules was associated with the development of Keynesian macroeconomics in the 1930s and extended into the growth theory of Harrod (1939), Domar (1946) and Solow (1956). The common formulation of these models held that all economic agents – consumers, producers and investors – select at any date t decision functions which depend only upon economic variables at date t. This gives rise, for example, to an aggregate consumption function such as $c(Y(t), r(t))$, which states that current aggregate consumption depends upon aggregate income $Y(t)$ and the interest rate $r(t)$. Although Keynes's writings demonstrate a deep understanding of the importance of expectations and other intertemporal considerations, the resulting macroeconomics which emerged was founded on entirely myopic decision rules. The two common explanations given to the formulation of myopic decision rules originate in issues of rationality and market imperfection. The first is a simple case of bounded rationality which results in extreme discounting of the future. The second explanation is based on the idea that Keynesian theory is not a theory of perfect competition and perfect price flexibility, rather, it must be interpreted as reflecting economic conditions in which price rigidity, rationing and quantity constraints are operative in various markets so that intertemporal substitutions are not generally feasible. This is particularly true for individuals with liquidity constraints. When such restrictions are operative, a consumer, a producer or an investor can respond only to contemporaneous variables and

cannot respond to future changes in prices or endowments.

Even through the period in which Keynesian macroeconomics provided the dominant intellectual tone, non-myopic models of behaviour were being developed. They became very influential through Ramsey (1928) on optimal intertemporal allocations, Modigliani and Brumberg (1954) on the life-cycle hypothesis and Friedman (1957) on the permanent income hypothesis. These contributions laid the foundations for modern thinking about intertemporal substitution and non-myopic decisions. In general, during the postwar period it was the study of consumer and investment behaviour which provided the arena for the debate: non-myopic decision models were derived mostly by microtheorists whereas both empirical researchers and macroeconomists tended to adopt more myopic decision rules. Apart from the Keynesian justification for myopic decision rules given above, two additional reasons were provided through this research, one empirical and the other conceptual. On the empirical side there is extensive evidence to suggest that individual consumption and investment are more sensitive to contemporaneous variables than would be implied by a non-myopic decision rule. On a more conceptual basis, an individual who wishes to make life-cycle plans must make his decision on the basis of his assessment of future events. Some of these events–like prices–might be observed on futures markets but others like future envowments, transfer payments or technology are uncertain and certainly unobservable by an economic analyst. Hence without some information about the non-observable conjectures of the decision maker about future events, it is not immediately clear how to identify empirically a consumer who follows non-myopic decision rules. In this connection, the 'permanent income hypothesis' may be viewed as a synthetic procedure which integrates the non-myopic nature of the consumer with his relative uncertainty about different components of his wealth (i.e. permanent versus transitory components of income).

In modern times, a further classification was made within the group of myopic decision making consumers by identifying the set of variables over which their utility function is defined. On the one hand there are the 'strict' life-cycle consumers whose utility extends only over consumption vectors consumed by them during their own life. On the other hand, when consumer interdependence is recognized with the extended family, non-myopic behaviour may be extended to allocations over present and future generations. Frequently the strict life-cycle hypothesis is modified by adding the total value of 'bequests' to the set of commodities over which the utility function is defined. Thus the utility function is written as $U(c(t), c(t+1), \ldots, c(t+T), B)$, where B is the value of bequests. Such extensions intend to accommodate an individual's concern for his extended family but is an unsatisfactory device. The fault of this formulation is seen from the fact that the decision rule which it implies is non-myopic over the life of the individual but entirely myopic with respect to dates beyond that. This myopic takes the form of insensitivity of this decision rule to future commodity prices, interest rates, endowments or technology.

Myopic Decision Rules and Intertemporal Consistency

When a non-myopic plan is formulated at time t, the decision maker will take into account all relevant variables $x(t), x(t+1), \ldots, x(t+T)$) and make a plan which will call for actions $(d(t), d(t+1), \ldots, d(t+T))$ to be taken at t and all subsequent dates up to time $t+T$. But now when date $t+\tau$ arrives will he carry out his plan $d(t+\tau)$ for this date? A consumer that would carry out his plans for dates $t+\tau$ for all $1 \leq \tau \leq T$ is said to be *intertemporally consistent*. In the original paper which raised this question, Strotz (1956) argued that, in general, consumers may not be intertemporally consistent. Alternatively, it was noted by Pollak (1968) that an optimizing individual could take into account future deviations from an initially chosen plan as a further constraint upon the set of feasible plans. For this reason a *sophisticated* planner was viewed by

Pollak as an individual who takes into account these restrictions whereas a *naive* planner ignores them. It is clear that even a naive planner always carries out his plans for the first date of the plan. These distinctions lead to the identification of three types of allocations: those planned (and carried out) by a sophisticated planner, those planned by a naive planner and those actually carried out by a naive agent. Both Strotz and Pollak assumed utilities to be additively separable and the discount function to be stationary in the sense that it may be written in the form $\delta(s-t)$ where t is the decision date, s is the date of consumption and $s \geq t$ with $\delta(0) = 1$. Under these conditions, the known theorem is that consistency of plans is equivalent to the condition $\delta(t) = e^{-\delta t}$ where δ is a constant. In this case, all three types of allocations specified above are the same.

Blackorby et al. (1973) extend this concept and analysis to the more general case of nonseparable preferences. Furthermore, Hammond (1976) formally synthesizes this literature on exogenously-changing tastes with a related literature on endogenously-changing tastes. Both articles formally demonstrate, for finitely-lived consumers, that inconsistency may arise only if at any time t the individual's ordering of consumption sequences which are identical up to time $t + \tau$ (but diverge thereafter) is different from the ordering of that individual at time $t + \tau$. Such preferences are, quite naturally, termed *inconsistent.* In general, consistent preferences give rise to intertemporally consistent plans for all possible budget constraints whereas inconsistent preferences induce consistent plans only in some special cases. Furthermore, unless preferences are consistent, neither naive nor sophisticated behaviour will, in general maximize any preference ordering nor even satisfy the weak axiom of revealed preference.

Intergenerational Equilibrium and the Neutrality Theory

The standard intergenerational allocation problem is formulated by specifying a sequence of generations each with its own endowment and preference. The crucial aspect of the problem is the interdependence in utilities where the utility of generation t depends upon the consumption, and, perhaps, the welfare level of generations of later dates. As a result of this interdependence, resources flow from generation t to generation $t + 1$. This structure is analogous to the intertemporal planning problem discussed earlier since it is natural to think of an individual at different points in time as a different generation of an infinitely lived extended family. However, it is also clear that the concept of *consistent* planning highlights a fundamental flaw in models of the family as an individually rational agent, namely, that future generations will likely reconsider the consumption plan selected by their ancestors. Consequently, for an allocation to be regarded as a possible social outcome, it must satisfy this elementary consistency requirement among the individual decision rules. Since the model of the infinitely lived rational individual may not satisfy this condition, a different conceptual foundation must be introduced. The superior framework which corrects this flaw views the family as a sequence of players in a noncooperative game. In an intergenerational equilibrium, each generation-player selects an optimal strategy of consumption and capital transfers given the strategies of the others. An equilibrium which is subgame perfect (see Selten, 1975) calls for strategies which satisfy the desired intergenerational consistency property. In this context the non-myopic decision rules are, in fact, non-myopic *strategies.*

The concept of intergenerational equilibrium was first proposed by Phelps and Pollak (1968) who studied it in the context of a simple aggregative model where the allowable strategies are savings functions which are linear in income and where preferences are additively separable. Subsequent contributions considered more general economies with broader strategy spaces. Although the aim of this research has been to provide a general theorem for the existence and characterization of perfect equilibrium (and thus a consistent plan) no such theorem has yet been proved. Significant progress has recently been achieved by Harris (1985).

The most controversial application of the theory of intergenerational equilibrium has been in the area of public policy. The Ricardian Equivalence theory holds that equilibrium utility allocations are invariant to changes in the method used to finance public expenditure. The idea of this theory is that no matter whether the public sector is financed by taxes or debt, the private allocation will be rearranged to neutralize any effect of public finance. This doctrine was recently reexamined by Barro (1974) in a formal model of an intergenerational equilibrium which postulated that preferences have a recursive representation. This means that the utility level of a member of generation t depends upon his own consumption and the utility level of his children, members of generation $t + 1$. Recursive representation is an important special case but such a representation exists only under very specialized conditions which are close to the concept of 'consistent preferences' discussed earlier. For this case, Barro was able to show that the existence of national debt had no real effect on the economy since for every specified method of public finance (debt versus taxes) there exists an intergenerational Nash equilibrium in which the utility allocation is the same as the utility distribution in the equilibrium without any public debt.

The proposed neutrality of public policy under non-myopic equilibrium strategies has dramatic consequences. It contrasts sharply with contemporary views that larger internal debts cause interest rates to increase. It is clear that these views are consistent with the theory which proposes that owing to intertemporal constraints individuals adopt myopic decision rules. Under myopic rules, increased internal debts would, in fact, cause interest rates to increase and private investments to be crowded out.

See Also

- ▶ Consumer Expenditure
- ▶ Intertemporal Equilibrium and Efficiency
- ▶ Ricardian Equivalence Theorem
- ▶ Time Preference
- ▶ Uncertainty and General Equilibrium

Bibliography

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Blackorby, C., D. Nissen, D. Primont, and R.R. Russell. 1973. Consistent intertemporal decision making. *Review of Economic Studies* 40: 239–248.

Von Böhm-Bawerk, E. 1889. *The positive theory of capital*. Trans. William Smart. London: Macmillan, 1891.

Domar, E.D. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14: 137–147.

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

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

Hammond, P.J. 1976. Changing tastes and coherent dynamic choice. *Review of Economic Studies* 43: 159–173.

Harris, C. 1985. Existence and characterization of perfect equilibrium in games of perfect information. *Econometrica* 53: 613–628.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Keynes, J. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: an interpretation of cross-section data. In *Post-Keynesian economics*, ed. K. Kurihara. New Brunswick: Rutgers University Press.

Phelps, E.S., and R.A. Pollak. 1968. On second-best national saving and game-equilibrium growth. *Review of Economic Studies* 35: 185–199.

Pollak, R. 1968. Consistent planning. *Review of Economic Studies* 35: 201–208.

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

Selten, R. 1975. Reexamination of the perfectness concept of equilibrium points in extensive games. *International Journal of Game Theory* 4: 25–55.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.

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

Myrdal, Gunnar (1898–1987)

Paul Streeten

Keywords

Cumulative causation; Economic development; Ex ante and ex post; Expectations;

Myrdal, G.; Non-economic variables; Planning; Population policy; Racial disadvantage; Stagflation; Underdevelopment

JEL Classifications
B31

Gunnar Myrdal was born in the province of Dalarná in Sweden. He attributed his faith in Puritan ethics and his egalitarianism to his sturdy farming background.

He was a student of the giant figures Knut Wicksell, David Davidson, Eli Heckscher, Gösta Bagge and above all Gustav Cassel. His personal friendship was warmest with Cassel, to whose chair in Political Economy at Stockholm University he succeeded (1933–9).

At first a pure theorist, Myrdal's year in the United States as a Rockefeller Fellow, following the crash of 1929, turned his interests to political issues. On his return to Sweden from America he, with his wife Alva, became active in politics. In 1935 he became a Member of Parliament. Together, they pioneered modern population policy. His involvements in Swedish politics between 1931 and 1938 turned him from a theoretical economist into a political economist and what he himself describes as an institutionalist. In 1938 the Carnegie Corporation selected him for a major investigation of the Negro problem in America, a project which resulted in *An American Dilemma* (1944a).

He returned to Sweden in 1942 and for five years was again involved in political activities. He headed the committee that drafted the social democratic post-war programme. He returned to Parliament and became a member of the board of directors of the Swedish Bank, chairman of the Swedish Planning Commission, and Minister for Trade and Commerce (1945–7). As Minister he arranged for a highly controversial treaty with the Soviet Union and was also involved in controversy over the dismantling of wartime controls. In 1947 he became Executive Secretary of the United Nations Economic Commission for Europe, to which he recruited an outstandingly able team. After ten years with the Commission in Geneva he embarked on a ten-year study of development in Asia, the result of which was the monumental *Asian Drama* (1968). In 1973 he was awarded the Nobel Prize in economics jointly with Friedrich von Hayek.

Methodological questions occupied Myrdal's thought throughout his life. They were already present in the young Myrdal's *Political Element in the Development of Economic Theory* (1930, English edition 1953). It was under the influence of the remarkable Uppsala University philosopher Axel Hägerström that he had begun to question the wisdom of the economic establishment.

Myrdal's doctoral dissertation on price formation and economic change (1927) introduced expectations systematically into the analysis of prices, profits and changes in capital values. The microeconomic analysis focused on planning by the firm. Many of these ideas were used in his later macroeconomic work, including *Monetary Equilibrium* (1931, English expanded translation 1939).

Much confusion had been caused by the lack of distinction between anticipations and results. The concepts *ex ante* and *ex post* that Myrdal developed greatly clarified the discussion of savings, investment and income, and their effects on prices. In anticipation, intention and planning, savings can diverge from investment; after the event they must be identical, because the community can save only by accumulating real assets. It is the process by which anticipations *ex ante* are adjusted so as to bring about the bookkeeping identity *ex post* that explains unexpected gains and losses as well as fluctuations in prices. Only in equilibrium are *ex ante* savings equal to *ex ante* investment, so that there is no tendency for prices to change. By introducing expectations into the analysis of economic processes he made a major contribution to liberalizing economics from static theory, in which the future is like the past, and to paving the way for dynamics, in which time, uncertainty and expectations enter in an essential way.

What is common to his three important later books, *The Political Element* (1930), *American Dilemma* (1944a) and *Asian Drama* (1968) is the emphasis on realistic and relevant research,

whether on economic problems, race relations, or world poverty, and with it the effort to purge economic thinking of systematic biases.

Starting on the study of Blacks in the United States, he soon discovered that he had to study 'American civilization in its entirety, though viewed in its implications for the most disadvantaged population group' (Introduction to *An American Dilemma*, Section 4). The way to reach objectivity was to state explicitly the value premisses of the study. These premisses were not chosen arbitrarily, but were what Myrdal called the 'American Creed' of justice, liberty and equality of opportunity. But while these value premisses were chosen for their relevance to American society, they corresponded to Myrdal's own valuations. The major contribution of the book, which Myrdal regarded as his war service, is the analysis of six decades after Reconstruction as a 'temporary interregnum' not a 'stable equilibrium', and of the incipient changes, on which the prediction of the Black revolt in the South was based.

Apart from his work on expectations and on racial problems, Myrdal is best known for his critique of conventional economic theory applied to underdeveloped countries.

Through his whole work run five lines of criticism of mainstream economic and social theory. First, his appeal for realism is not a critique of abstraction. His criticism is that irrelevant features are selected and relevant ones ignored ('opportunistic ignorance'). A second line of criticism has been the narrow or abstract definitions of development, economic growth, or welfare. The actual needs and valuations of people, not the abstractions of statisticians or the empty concepts of metaphysicians, should be the basis for formulating aims. His third line of criticism is directed at the narrow definitions and the limits of disciplines. The essence of the institutional approach, advocated by Myrdal, is to bring to bear all relevant knowledge and techniques on the analysis of a problem. In an interdependent social system there are no economic, political or social problems, there are only problems. His fourth line of criticism is directed at spurious objectivity which, under the pretence of scientific analysis, conceals political valuations and interests. Myrdal argues that this pseudoscience should be replaced by explicit valuations. He is, of course, aware of the complex nexus between valuations and facts but, ever since his youthful *Political Element,* has constantly fought the inheritance of natural law and utilitarianism, according to which we can derive recommendations from pure analysis. A fifth line of criticism is directed against biases and twisted terminology. He lays bare the opportunistic interests and the 'diplomacy' underlying the use of such concepts as 'United Nations', 'international', 'values', 'welfare', 'developing countries', 'unemployment', 'the free world'. The features against which these lines of criticism are advanced are combined in the technocrat. He isolates economic (or other technical) relations from their social context; he neglects social and political variables and thereby ministers to the vested interests that might otherwise be hurt; he pretends to scientific objectivity and is socially and culturally insensitive. Since the majority of experts, academics and planners are of this type, he has ruffled many feathers.

The question may be asked whether the narrow technocrat cannot be replaced by an approach that introduces social variables openly into the formal model?

Myrdal's answer would be, yes and no. In certain areas, a widening or redefinition of concepts can be successful. The productive effects of better nutrition can be studied and the line between investment and consumption be redrawn. The influence of climate, of attitudes, and of institutions can be introduced as constraints or as variables. An agricultural production function can be constructed in which health, education, distance from town, and so on figure as 'inputs'. 'Capital' can be redefined so as to cover anything on which expenditure of resources now raises the flow of output later.

But there are limits to such revisionism. These limits apply both to the analysis of facts and to recommendations of policies. On the factual side, the reformulation runs into difficulties if the connection between expenditure now and 'yield' later is only tenuous, as in the initiation of a birth control programme or a land reform.

In the analysis of values, the construction of a social welfare function is not, in Myrdal's view, a

logical task. The unity of a social programme of a party is unlike that of a computer program or a logically consistent system, and more like the unity of a personality. It is discovered not only by deductive reasoning but by empathy, imagination, and even artistic and intuitive understanding. Means and ends, targets and instruments, are misleading ways of grasping the valuations of a class, an interest group or a whole society, for their unity is not logical but psychological.

In *Asian Drama* the explicitly formulated valuations are the 'Modernization Ideals'. A list would include rationality, planning for the future, raising productivity, raising levels of living, social and economic equalization, improved institutions and attitudes, national consolidation, national independence, political democracy, social discipline.

An important idea in Myrdal's arsenal of ideas is that of circular or cumulative causation (or the vicious – or virtuous – circle), first fully developed in *An American Dilemma*. It postulates increasing returns through specialization and economies of scale and shows how small advantages are magnified.

The principle goes back to Wicksell who, in *Interest and Prices* (1898), had analysed divergences between the natural and the market rates of interest in terms of upward or downward cumulative processes, until the divergence was eliminated. Wicksell pointed out that, if banks keep their loan rate of interest below the real rate of return on capital, they will encourage expansion of production and investment in plant and equipment. As a result, prices will rise and will continue to rise cumulatively as long as the lending rate is kept below the real rate.

The principle of cumulative causation can be used to show movements away from an equilibrium position as a result of the interaction of several variables. Myrdal has not always been entirely clear in the formulation of this important principle, and there has been the suggestion that any form of circular or mutual causation or interaction is cumulative and hence disequilibrating. This would be false, for a series of mutually caused events can, after a disturbance, rapidly converge either on the initial or on some other point of stable equilibrium. In order to get instability, a cumulative movement away from the initial situation, the numerical values of the coefficients of interdependence have to be above a critical minimum size. For example, an increase in consumption will raise incomes which in turn will raise consumption, and so on, ad infinitum. But as long as the marginal propensity to consume is less than unity, the infinite series will converge on a finite value.

The notion of cumulative causation was applied by Myrdal most illuminatingly to price expectations (*Monetary Equilibrium*) and to the relations between regions (*Economic Theory and Underdeveloped Regions,* 1957; American title: *Rich Lands and Poor*). He showed how the advantages of growth poles can become cumulative, while the backward region may be relatively or even absolutely impoverished.

Myrdal applied the notion of sociological variables, such as the prejudices against Negroes and their level of performance (low skills, crime, disease, and so on); to economic variables; and, above all, to the interaction of so-called 'economic' and 'non-economic' variables. Thus, the relation between better nutrition, better health and better education, higher productivity and hence ability further to improve health, education and nutrition shows that the inclusion of non-economic variables in the analysis opens up the possibility of numerous cumulative processes to which conventional economic analysis is blind. It also guards against uni-causal explanations and panaceas.

The revolutionary character of the concept of cumulative causation is brought out by the fact that interaction takes place not only within a social system in which the various elements interact, but also in time, so that memory and expectations are of crucial importance. The responses to any given variable, say a price, are different according to what the history of this variable has been. It is this dynamic feature of analysis and its implications for policy that distinguishes Myrdal's approach from that of economists who think in terms of general equilibrium.

In *Economic Theory and Underdeveloped Regions* (1957), and later in *Asian Drama*

(1968), he used the concepts 'backwash' and 'spread' effects to analyse the movement of regions or whole countries at different stages of development and the effects of unification. It is a highly suggestive, realistic and fruitful alternative explanation to that of stable equilibrium analysis, usually based on competitive conditions and diminishing returns, and concluding that gains are widely and evenly distributed.

Like the Marxists, Myrdal emphasizes the unequal distribution of power and property as an obstacle not only to equity but also to efficiency and growth. But his conclusion is not Marxist. He regards a direct planning of institutions and shaping of attitudes (what Marx regarded as part of the superstructure) as necessary, though very difficult, partly because he believes that attitudes and institutions are inert, and partly because the policies which aim at reforming attitudes and institutions are themselves part of the social system, part of the power and property structure. There are clearly also logical difficulties in operating on variables that are thought to be fully determined within the system.

In *Asian Drama* Myrdal criticizes the kind of government he calls the 'soft state'. This critique has sometimes been misunderstood. It is plain that 'softness' in Myrdal's sense is quite compatible with a high degree of coercion, violence and cruelty. The Tamils in Sri Lanka, the Indians in Burma, the Chinese in Indonesia, the Hindus in Pakistan, the Moslems in India, the Biharis in Bangladesh – to take six states he calls 'soft' – could not claim excessively soft treatment. 'Soft states' also go in for military violence, both internal and external. Their 'softness' lies in their unwillingness to coerce in order to implement declared policy goals. It is not the result of gentleness or weakness but reflects the power structure and a gap between real intentions and professions.

Myrdal applied his method to the analysis of inflation combined with widespread unemployment in the developed countries of the West in the 1970s, and either coined or was one of the first to use the term 'stagflation'. He attributes the situation to the organization of producers as pressure groups, and the dispersion and comparative weakness of consumers, to the tax system which encourages speculative expenditures, to the structure of markets and to the methods of oligopoly administrative pricing, and he condemns inflation as a socially highly divisive force.

The approach favoured by Myrdal is one of neither Soviet authority and force nor of capitalist laissez-faire but of a third way: that of using prices for planning purposes and of attacking attitudes and institutions directly to make them the instruments of reform. His approach has more affinity with those socialists who were dismissed by Marx as utopian. The difficulty is that any instrument, even if used with the intention to reform, within a given power structure may serve the powerful and re-establish the old equilibrium. Even well-intentioned allocations, rationing, licensing and controls may reinforce monopoly and big business. How does one break out of this lock? Myrdal does not draw revolutionary conclusions but relies on the, admittedly difficult, possibility of self-reform that arises, in both the American Creed and in the Modernization Ideals, from the tensions between preferred and proclaimed beliefs and actions.

Both An *American Dilemma* and *Asian Drama* are books about the interaction and the conflict between ideals and reality, and about how, when the two conflict, one of them must give way. Much of conventional economic theory is a rationalization whose purpose it is to conceal that conflict. But it is bound to reassert itself sooner or later. When this happens, either the ideals will be scaled down to conform to the reality or the reality will be shaped by the ideals.

See Also

- ▶ Ex Ante and Ex Post
- ▶ Institutional Economics

Selected Works

1927. *Prisbildningsproblemet och föränderligheten.* Uppsala: Almqvist & Wiksell.

1930. *Vetenskap och politik i nationalokonomien*. Stockholm: P.A. Norstedt.

1931. Om penningteoretisk jämvikt. *Ekonomisk Tidskrift* 33, 191–302.
1932. *Das Politische Element in der nationalökonomischen Doktrinbildung*. Berlin: Junker and Dünnhaupt.
1933a. *The cost of living in Sweden 1830–1930.* London: P.S. King & Son.
1933b. Der Gleichgewichtsbegriff als Instrument der geldtheoretischen Analyse. In *Beiträge zur Geldtheorie,* ed. P. Friedrich von Hayek. Vienna: Springer-Verlag (expanded version of Myrdal, 1931).
1933c. *Konjunktur och offentlig hushållning. Bihang til riksdagens protokoll*, 1st collection, Appendix III.
1933d. Das Zweck-Mittel-Denken in der Nationalökonomie. *Zeitschrift für Nationalökonomie* 4, 305–29.
1934a. *Finanspolitikens e konomiska verkningar*. Stockholm: P.A. Norstedt.
1934b. (With A. Myrdal.) *Kris i befolkningsfrågan.* Stockholm: A. Bonnier.
1939. *Monetary equilibrium.* London: Hodge.
1940. *Population: A problem for democracy.* Cambridge, MA: Harvard University Press.
1944a. *An American dilemma: The Negro problem and modern democracy.* New York: Harper. Paperback edn, New York: McGraw-Hill, 1964.
1944b.*Varning för fredsoptimism.* Stockholm: Bonniers.
1945. *Warnung gegen Friedensoptimismus*. Zurich: Europa Verlag.
1953. *The political element in the development of economic theory.* Trans. P. Streeten. London: Routledge & Kegan Paul. Cambridge, MA: Harvard University Press. (Originally published in German.)
1955. *Realities and illusions in regard to intergovernmental organisations*. Oxford: Oxford University Press.
1956a. *Development and underdevelopment: A note on the mechanism of national and international inequality.* Cairo: National Bank of Egypt.
1956b. *An international economy: Problems and prospects*. London: Routledge & Kegan Paul.
1957. *Economic theory and underdeveloped regions.* London: Duckworth; New York: Harper.
1958. *Value in social theory: A selection of essays on methodology,* ed. P. Streeten. London: Routledge & Kegan Paul.
1960. *Beyond the welfare state.* New Haven: Yale University Press.
1961. Value-loaded concepts. In *Money, growth and methodology and other essays in honor of Johan Åkerman,* ed. H. Hegeland. Lund: Gleerup.
1963. *Challenge to affluence.* New York: Pantheon.
1968. *Asian drama: An inquiry into the poverty of nations*. New York: Twentieth Century Fund.
1969. *Objectivity in social research*. New York: Pantheon.
1970a. *The challenge of world poverty: A world anti-poverty program in outline*. New York: Pantheon.
1970b. The 'soft state' in underdeveloped countries. In *Unfashionable economics,* ed. P. Streeten. London: Weidenfeld & Nicolson.
1972. *Against the stream: Critical essays on economics*. New York: Pantheon.

Bibliography

Myrdal's intellectual development is described in J. Angresano, *The Political Economy of Gunnar Myrdal,* Cheltenham: Edward Elgar, 1997. The influence of *An American Dilemma* on public policy has been sufficiently great to generate scholarly studies on its genesis and impact, including W. Jackson, *Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987,* Durham: University of North Carolina Press, 1990, and D. Southern, *Gunnar Myrdal and Black–White Relations: The Use and Abuse of an American Dilemma, 1944–1969,* Baton Rouge: Louisiana State University Press, 1987.