B

Babbage, Charles (1791–1871)

Maxine Berg

Charles Babbage is rarely regarded as a major contributor to economic thought. His name is synonymous with the early origins of the computer, and he was an important figure in early nineteenth-century scientific circles. He was educated at Trinity College and Peterhouse, Cambridge, and while still a student started the Analytical Society with Herschel and Peacock, for reforming mathematics in Britain. His interest in mathematics was the foundation for his later contributions to science, economics and statistics. After Cambridge, Babbage moved to London, where he began his lifelong work on his analytical engine and became a leading participant in scientific circles. He joined the Royal Society and was a founding member of the Cambridge Philosophical Society and the Royal Astronomical Society. Later he was to be one of Newton's illustrious successors in the Lucasian Chair of Mathematics at Cambridge. But he was also a radical if maverick intellectual and political critic. He wanted to see science reformed, to see British science play a leading part in theoretical advance, and to see this science related closely to applied technology. He also demanded a role for the state in providing support for science and university education, and for establishing a policy on technology. He wrote a controversial attack on the Royal Society, *Reflections on the Decline of Science and Some of its Causes* (1830), and was one of the founding trustees of the British Association for the Advancement of Science, with the purpose of bringing science and technology, from the provinces as well as the metropolis, into the forefront of culture and society.

Babbage was an early promoter of industrial exhibitions as a part of meetings of the British Association; he participated in the Mechanics Section of the Association and later wrote a book on the Great Exhibition of 1851. He took part in the great controversies over religion and science in the period, and wrote the *Ninth Bridgewater Treatise* in 1837 (2nd edn 1838), conveying his belief in a Newtonian universe, with a scientific Deity.

Politically, Babbage was a liberal Whig; he chaired an election committee and stood twice for Finsbury. He denounced election corruption and bribery, attacked church preferments and tithes, and was a firm supporter of the Reform Bill. His political pamphlet on income tax showed a concept of moderate reform. He identified an electoral system based on one man, one vote with the 'advance of socialism', for where the poor were in the majority they would vote for low taxes for themselves and high taxes for the rich, 'thus destroying private enterprise'.

Babbage's social and academic context was clearly that of early nineteenth-century liberal-scientific circles, and he participated in the salon

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

culture of the day. But there was another very important component to his intellectual make-up: an abiding interest in practical mechanics and a fascination with contemporary industrial technology. He learned from manufacturers, large and small, mechanical engineers and above all from the skilled artisans he never ceased to praise. Developing the analytical engine was itself a task of scientific and mathematical reasoning combined with practical invention. The continental tour he made in 1827–1828, which was to be so formative to his later work, was not in the company of a scientific friend or even a servant, but with one of the artisans who had worked on the building of the analytical engine. Travelling through the Low Countries, Germany, Austria and Italy with a prolonged stay in Naples, Babbage lost no opportunity to visit local workshops and factories.

His transcendence of contemporary social and intellectual boundaries was the real basis for his brilliant and utterly original foray into political economy. *On the Economy of Machinery and Manufactures* (1832) was immensely popular: there were four editions in 2.5 years, it was reprinted in the United States and translated into four continental languages. Babbage wanted to present his readers with the mechanical principles of arts and manufactures, and he hoped also to be read by the intelligent working man. To this extent the book fell within the contemporary genre of industrial-technological literature; indeed, part of it had been published in 1829 as a part of the *Encyclopedia Metropolitana*. Tracts on the steam engine, histories of the cotton industry and industrial manuals, dictionaries and encyclopedias were very popular at the time. Andrew Ure's later *Philosophy of Manufactures* (1835), an extraordinary panegyric on the factory system and steam-powered machinery, was very much a product of this genre, but it completely lacked the analysis of Babbage's contribution. The latter was much more than popular industrial observation. It was an analysis based on economic principles, especially the Smithian account of the division of labour, of manufacturing technology and the organization of industrial work. Babbage's obvious first-hand knowledge of a wide variety of industrial and business processes, combined with general analysis of production systems, made the work a tour de force. At a time of anxiety and ambiguity over the reception of new technology, he also offered authoritative policy statements on a wide range of machinery issues including patent reform, export of machinery, crises of overproduction, and technological unemployment.

The book's intellectual situation in relation to political economy was not, however, easily apparent, and apart from Mill and Marx few appreciated its significance to their discipline. Before he wrote the book Babbage had intended to deliver a series of lectures in Cambridge on the Political Economy of Manufactures, but this never materialized. He himself conceded that his first edition did not profess to examine questions of political economy, and he attempted to correct this in the next edition by introducing three new chapters: 'The new system of manufactures', 'The effects of machinery in reducing the demand for labour', and 'On money as a medium of exchange'. But most of the topics raised by Babbage were also foreign to contemporary classical political economy. Moving back to Smith, he analysed industrial organization and the microeconomics of the manufacturing firm, never losing sight of technological constraints and opportunities.

The book was initially criticized for failing to give due attention to the factory system and steam-powered textile technology. But this was precisely its strength, for it analysed the factory and the workshop as parts of the more general organization of work, and examined machinery in the context of a more general discussion of technology, including skill. Babbage's close observation of skills and hand processes as well as machinery, of the workshop as well as the factory, was anyway a more accurate perception of contemporary industrial practice than a work concentrating on the outstanding and atypical phenomenon of the factory would have been.

> Babbage analysed what he called 'the domestic economy of the factory'. He sought to specify what arrangement of production would succeed in selling articles at a minimum price, and he made a careful analysis of economies of scale in relation to

> the division of labour, distinguishing the dynamics of the factory from those of the workshop. He developed Smith's principle of the division of labour to a further refinement, introducing the significance of the division of skill, or the division of mental and manual labour. Vital, he believed, to the success of any organization of work was his 'Babbage Principle':
>
> that the master manufacturer by dividing the work to be executed into different processes, each requiring different degrees of skill or of force, can purchase exactly that precise quantity of both which is necessary for each process.

From this emphasis on the economy of skill, Babbage introduced novel discussions of the role of accounting, time and motion studies, communications innovations, and an analysis of machine functions. He was particularly concerned with the significance of precision and measurement in all processes, with the regularity of production, and with the planning of layout. He thus regarded as some of the greatest innovations not the celebrated power techniques themselves, but the processes which helped to make the new machines work properly, for example the steam engine governor and lubrication or grease. His interest in measurement led him to support all manner of instruments for counting machines and human actions; and he devised a detailed questionnaire as a basis for job studies and early time and motion studies. He also analysed as had no one before him the role of the speed of production and the intensity of labour in increasing output. Introducing machinery was only one incomplete route to increasing productivity; the productivity of labour could be rapidly improved through greater order, precision and labour discipline. Babbage noticed the convergence of technological and economic principles on topics such as velocity and copying; in a long discussion of the significance of copying techniques he pointed out the parallels between printing, casting and moulding, stamping and turning.

This core analysis of workshop organization was complemented by topical commentary on profit sharing, technological unemployment and trade unions. An important radical departure on wages and labour was provided in his 'New System of Manufactures' which argued for a piecerate wage system and profit sharing, if not cooperation, as the key to overcoming the long-standing worker opposition to machinery. This was the problem which Babbage along with many of his contemporaries believed to be the major brake on Britain's industrial progress. The system was a far-reaching proposal for a worker's stake in increasing productivity, for collective decision-making on hiring, dismissal and the organization of the works. Where the system prevailed, modern methods would be chosen and an extensive division of labour introduced, not to control and subordinate labour, but as a cooperative decision by workers for the most efficient methods. The lengths to which his suggestions went were probably surpassed only by radical and Owenite cooperatives. When it came to practical implementation Babbage held out little hope of any appeal of the system to large established firms, but thought that groups of artisans and small firms would lead the way. Babbage's chapters on trade unions and machinery and employment were, however, the comments of a reformer not a radical. He attacked the truck system, but warned that trade unions could well lead to more rapid displacement of labour through machinery or industrial relocation. Dealing with technological underemployment, he used the case of hand-weaving and the power loom, arguing that the only solution lay in better workers' planning through such institutions as savings banks and friendly societies.

Babbage's use of practical observation and statistical data, and his critique of political economy's 'closest philosophers', induced him, with Richard Jones, J.E. Drinkwater, Malthus and Quetelet, to form a Statistics Section of the British Association in 1833, followed later by the Statistical Society of London. The Statistical Section, Section F, was confined to the presentation of statistical data, avoiding areas of political controversy. But the less restrictive London Society made its brief the connection of political economy to the statistical investigation of economic improvement. Babbage was the first President of Section F, and wrote several statistical papers. His earlier 'Letter to the Right Hon T.P. Courtenay on the proportional number of births of the two sexes under different circumstances' (1829) compared the demographic structures of the Kingdom of

Naples, France, Prussia and Westphalia. Much later he wrote 'On the statistics of lighthouses' for the Brussels Congress of Statistics in 1853, and 'The clearing house', read to the London Statistical Society and printed in its memoirs in 1856. Babbage also wrote a book on insurance, *A Comparative View of the Various Institutions for the Assurance of Lives* (1826), and is remembered for his revised actuarial tables and his popular presentation of a difficult subject.

Babbage certainly produced an original and far-seeing economic analysis of industry in *On the Economy of Machinery and Manufactures*. He applied the principles of the division of labour he elaborated to his perception of the sciences. The ultimate result of the division of skills and especially the mental division of labour was the 'science of calculation'. He argued that the science of calculation, like any technology, would be developed to a degree where machinery would take over all numerical calculation. Arithmetical exercise would thus be separated from mathematical reasoning, and the 'science of calculation' harnessed to the analytical engine would become the science of all sciences. Babbage's ultimate vision for Britain's industrial progress was one of a computer-run technology.

Selected Works

1830. *Reflections on the decline of science and some of its causes*. London: B. Fellowes.
1832. *On the economy of machinery and manufactures*. London: Charles Knight. 4th ed., 1835.
1838. *The ninth bridgewater treatise*, 2nd ed. London: John Murray.
1851. *The exposition of 1851*. London: John Murray.
1864. *Passages from the life of a philosopher*. London: Longman & Co.

References

Dubbey, J.M. 1978. *The mathematical work of Charles Babbage*. Cambridge: Cambridge University Press.
Hyman, A. 1982. *Charles Babbage, pioneer of the computer*. Oxford: Oxford University Press.
Rosenberg, Nathan. 1994. Charles Babbage: Pioneer economist, chapter 2. In *Exploring the black box: Technology, economics, and history*. Cambridge: Cambridge University Press.
Rosenberg, Nathan. 2000. Charles Babbage in a complex world. In *Complexity and the history of economic thought*, ed. D. Colander. London: Routledge.
Swade, Doron. 2001. *The difference engine: Charles Babbage and the quest to build the first computer*. New York: Viking Press.

Babeuf, François Noël (1764–1797)

A. Courtois

François Noel Babeuf, called *Caius Gracchus*, was born in Saint Quentin in 1764 and died at Vendôme on 24 February 1797. Left to his own resources at the age of sixteen, his youth was stormy, and his whole life wild and irregular. From the commencement of the Revolution he wrote in the journal *Le correspondant Picard*, articles so violent in tone that he was brought to trial. His acquittal, 14th July 1790, did little to calm him. Appointed administrator of the Département of the Somme, he soon had to be dismissed from the office. This was the time at which he took the name of *Caius Gracchus*, posing as a *Tribun du peuple*. He gave the same name to a journal, which he had previously carried on under the sub-title of *Défenseur de la liberté de la presse*. All this took place shortly after the fall of Robespierre from power. This for a time had his approval; but he soon returned to his earlier views and appealed to those violent passions which, as a demagogue, he knew how to rouse. He gathered round him, under the name of the *Secte des Egaux*, all the old Montagnards who were dissatisfied with the régime of the Thermidorians. The object of this sect, which drew its inspiration from some of the sentimental ideas of J.J. Rousseau, was to destroy inequality of condition, with the object of attaining the general good. Sylvain Maréchal, author of a *Dictionnaire des Athées*, Buonarroti, who claimed to be descended from Michael Angelo, with Amand and Antonelle, who did

not, it is true, remain associated long, and some others, formed the staff which recognised Babeuf as their chief. Working with feverish activity, they gathered round them a considerable number of adherents. The place where their club met was the Pantheon. At first orderly, their meetings became tumultuous and threatening and were prolonged far into the night. Attending armed, they prepared to resist by force the dissolution of the club which the authorities had determined on. General Bonaparte, acting with much tact, contrived to close the meetings of the club, but the members formed themselves forthwith into a secret society, and gradually, by winning over soldiers and police, became a formidable body, numbering nearly 17,000 able-bodied and armed men, without including the Faubourgs Saint-Antoine and Saint-Marceau, which were at their back. Addressing themselves to the masses, they published a manifesto written by Sylvain Maréchal in his most inflammatory style.

> We desire [said they] real equality or death. This is what we want. And we will have real equality, no matter what it costs. Woe to those who come between us and our wishes. Woe to him who resists a desire so resolutely insisted on. . .. If it is needful, let all civilization perish, provided that we obtain real equality. . .. The common good, or the community of goods. No further private property in land; the land belongs to no private person. We claim, we require the enjoyment of the fruits of the land for all; the fruits belong to the whole world [etc].

Instructions in great detail as to the methods of raising insurrectionary movements were added.

> Those who hinder us shall be exterminated; . . . shall all alike be put to death: Those who oppose us or gather forces against us; strangers, of whatever nation they may be, who are found in the streets; all the presidents, secretaries, and officers of the royalist (*sic*) conspiracy of Vendémiaire, who may also dare to show themselves.

If the lives of men were to be treated thus, one may guess what fate was reserved for their property. But, after massacres and spoliations, what was to come of it all? The public authorities were to organise employment; there was to be only one source of employment, the state, with subdivisions devised to meet the wants, somewhat rudimentary, of the community. Every one was to have a right to lodging, clothes, washing, warming, and lighting, to food, *médiocre mais frugale*, to medical attendance. This is much what Louis Blanc, who appears to have sought his inspiration among the decrees of the *République des Egaux*, enunciated in more methodical and sober language. 'Every one is to work as he is able, and to consume according to his wants.'

The secret was well kept; it was only a few hours before the moment fixed for the explosion of the conspiracy (May 1796) that a captain, named Grisel, revealed it to the directory. Decisive steps were taken at once; a vigorous watch was kept, while the public authorities seized the leaders and their papers.

Babeuf and Darthé, condemned to death the 23rd of February 1797, stabbed themselves before the tribunal. Life still lingering on, they were guillotined the next day. Buonarroti and Sylvain Maréchal, condemned to exile (*déportation*), died, the first in 1837, the second in 1803.

It may be added that Babeuf seems to have had rather a disordered brain than an absolutely criminal disposition. He died with courage, leaving his wife a written paper declaring his conviction that he had always been a 'perfectly virtuous man'.

Bachelier, Louis (1870–1946)

Benott B. Mandelbrot

Keywords

Bachelier, L.; Brownian motion; Efficient markets; Levy, P.; Martingales; Poincaré, H.; Probability; Random walk model; Speculation; Wiener process

JEL Classifications

B31

Bachelier was born in Le Havre, France, on 11 March 1870 and died in Saint-Servan-sur-Mer, Ille-et-Vilaine, on 28 April 1946. He taught at Besançon, Dijon and Rennes and was professor at Besançon from 1927 to 1937.

The unrecognized genius is one of the stock figures of popular history, and it is also a platitude of which many examples dissolve upon careful examination. But the story of Louis Bachelier is in perfect conformity to all the clichés. He invented efficient markets in 1900, 60 years before the idea came into vogue. He described the random walk model of prices, ordinary diffusion of probability – also called Brownian motion – and martingales, which are the mathematical expression of efficient markets. He even attempted an empirical verification. But he remained a shadowy presence until 1960 or so, when his major work was revived in English translation.

This major work was his doctoral dissertation in the mathematical sciences, defended in Paris on 19 March 1900. Things went badly from the start: the committee failed to give it the 'mention très honorable', key to a university career. It was very late, after repeated failures, that Bachelier was appointed to the tiny University of Besançon. After he had retired, the university archives were accidentally set on fire and no record survives, not even one photograph. Here are a few scraps I have managed to put together.

We begin with the proverbial episode of the grain of sand, or the lack of a nail. Bachelier made a mathematical error that is recounted in a letter the great probabilist Paul Levy wrote me on 25 January 1964:

> I first heard of him around 1928. He was a candidate for a professorship at the University of Dijon. Gevrey, who was teaching there, came to ask my opinion. In a work published in 1913, Bachelier had defined Wiener's function (prior to Wiener) as follows: In each interval $[n\tau, (n+1)\tau]$, he considered a function $X(t|\tau)$ that has a constant derivative equal to either + v or - v, the two values being equiprobable. He then proceeded to the limit $\tau \to 0$, keeping v constant, and claimed he was obtaining a proper function $X(t)$! Gevrey was scandalized by this error. I agreed with him and Bachelier was blackballed.

I had forgotten it when in 1931, reading Kolmogorov's fundamental paper, I came to 'der Bacheliers Fall'. I looked up Bachelier's works, and saw that this error, which is repeated everywhere, does not prevent him from obtaining results that would have been correct if only he had written $v = C\tau^{-1/2}$, and that, prior to Einstein [1905] and prior to Wiener [circa 1925], he has seen some important properties of the Wiener function, namely, the diffusion equation and the distribution of $\max_{0 < \tau < t} X(\tau)$.

We became reconciled. I had written to him that I regretted that an impression, produced by a single initial error, should have kept me from going on with my reading of a work in which there were so many interesting ideas. He replied with a long letter in which he expressed great enthusiasm for research.

That Levy should have played this role is tragic, for his own career also nearly foundered because his papers were not sufficiently rigorous for the mathematical extremists.

The second and deeper reason for Bachelier's career problems was the topic of his dissertation: 'Mathematical theory of speculation' – not of (philosophical) speculation on the nature of chance, rather of (money-grubbing) speculation on the ups and downs of the market for consolidated state bonds: '*la rente*'. The function $X(t)$ mentioned by Levy stood for the price of *la rente* at time t. Hence, the delicately understated comment by Henri Poincaré, who wrote the official report on this dissertation, that ' the topic is somewhat remote from those our candidates are in the habit of treating'. One may wonder why Bachelier asked for the judgement of unwilling mathematicians (assigning a thesis subject was totally foreign to French professors of that period), but he had no choice: his lower degree was in mathematics and probability was taught by Poincaré.

Bachelier's tragedy was to be a man of the past and of the future but not of his present. He was a man of the past because gambling is the historical root of probability theory; he introduced the continuous-time gambling on *La Bourse.* He was a man of the future, both in mathematics (witness the above letter by Levy) and in economics. Unfortunately, no organized scientific community of his time was in a position to understand and

welcome him. To gain acceptance for himself would have required political skills that he did not possess, and one wonders where he could have gained acceptance for his thoughts.

Poincaré's report on the 1908 dissertation deserves further excerpting:

> The manner in which the candidate obtains the law of Gauss is most original, and all the more interesting as the same reasoning might, with a few changes, be extended to the theory of errors. He develops this in a chapter which might at first seem strange, for he titles it 'Radiation of Probability'. In effect, the author resorts to a comparison with the analytical theory of the propagation of heat. A little reflection shows that the analogy is real and the comparison legitimate. Fourier's reasoning is applicable almost without change to this problem, which is so different from that for which it had been created. It is regrettable that [the author] did not develop this part of his thesis further.

While Poincaré had seen that Bachelier had advanced to the threshold of a general theory of diffusion, he was notorious for lapses of memory. A few years later, he took an active part in discussions concerning Brownian diffusion, but had forgotten Bachelier.

Comments in a *Notice* Bachelier wrote in 1921 are worth summarizing:

> 1906: *Théorie des probabilités continues.* This theory has no relation whatsoever with the theory of geometric probability, whose scope is very limited. This is a science of another level of difficulty and generality than the calculus of probability. Conception, analysis, method, everything in it is new. 1913: *Probabilités cinématiques et dynamiques.* These applications of probability to mechanics are the author's own, absolutely. He took the original idea from no one; no work of the same kind has ever been performed. Conception, method, results, everything is new.

The hapless authors of academic *Notices* are not called upon to be modest, but Louis Bachelier had no reason for being modest. Does anyone know more about him?

See Also

► Wiener Process

Selected Works

1900. Théorie de la spéculation. *Annales de l'Ecole normale supérieure,* 3rd series 17: 21–86, trans. A. Boness. In: *The random character of stock market prices,* ed. P.H. Cootner. Cambridge, MA: MIT Press, 1967.

1901. Théorie mathématique des jeux. *Annales de l'Ecole normale supérieure,* 3rd series 18: 143–210.

1906. Théorie des probabilités continues. *Journal des Mathématiques Pures et Appliquées,* 6th series 2: 259–327.

1910a. Les probabilités à plusieurs variables. *Annales de l'Ecole normale supérieure,* 3rd series 27: 340–360.

1910b. Mouvement d'un point ou d'un système soumis à l'action des forces dépendant du hasard. *Comptes rendus de l'Académie des sciences* 151: 852–855.

1912. *Calcul des probabilités.* Paris: Gauthier-Villars.

1913. Les probabilités cinématiques et dynamiques. *Annales de l'Ecole normale supérieure,* 3rd series 30: 77–119.

1924. *Le jeu, la chance et le hasard*. Paris: E. Flammarion.

1937. *Les lois des grands nombres du calcul des probabilités.* Paris: Gauthier-Villars.

1938. *La spéculation et le calcul des probabilités.* Paris: Gauthier-Villars.

1939. *Les nouvelles méthodes du calcul des probabilités.* Paris: Gauthier-Villars.

Backwardation

Masahiro Kawai

Using the language of the London Stock Exchange, 'backwardation' is a fee paid by a seller of stocks (or securities) to the buyer for the privilege of deferring delivery of them. Hence it means that the futures price (i.e. the current price for the future delivery) falls short of the spot price

(i.e. the current price of immediate delivery). 'Contango', the reverse of backwardation, is a fee paid by the buyer who wants to postpone delivery, and means that the futures price exceeds the spot price. These terms may be extended to any futures transaction.

Keynes (1923, pp. 255–66, 1930, ch. 29) and Hicks (1946, pp. 130–40) advanced the theory of 'normal' backwardation; namely, the situation where the futures price of commodities is a downwardly biased prediction of the spot price at delivery time. Since normal backwardation is tantamount to the presence of a positive risk premium, hedgers as a whole take a short futures position of the commodities, and speculators as a group a long position. The theory of normal backwardation attempts to explain why hedgers tend to go short in futures.

Keynes and Hicks explained the existence of normal backwardation on technological grounds. That is, technological conditions in production and consumption (including demand activities by manufacturers who use the commodities as inputs) are such that producers must look much further ahead than consumers, because the former may already have committed themselves to production while the latter have a freer hand about acquiring the commodities. Thus there exists a greater desire to cover planned production (supplies) than to cover planned consumption (demands), and hedgers as a whole have a tendency to go short in futures. In order to persuade speculators to assume a matching long position, a positive risk premium has to be offered, hence a 'normal' backwardation.

Although this technological explanation is valid for typical commodity markets, it does not apply to all markets. Consider the following equilibrium conditions in the spot and futures markets at time 0:

$$Q_{0,0} + Z_{-1} = C_{0,0} + K_0 \quad \text{(Spot Market Equilibrium)}$$

$$Q_{0,1} + K_0 - C_{0,1} = Z_0 \quad \text{(Futures Market Equilibrium)}.$$

The variables Q, C, K and Z denote output supply, consumption, storage and futures speculation, respectively. The subscripts signify time; $Q_{t,s}$ (or $C_{t,s}$) is output or consumption) planned at time t and actually supplied (or demanded) at time s, K_0 is the amount carried from time 0 to time 1, and Z_0 is the quantity of speculative futures contracts purchased (if $Z_0 > 0$, or sold if $Z_0 < 0$) at time 0 for time 1 delivery. In the case of typical commodities, Q, C, $K > 0$ and $Z \lessgtr 0$. (For more detailed discussions about the market equilibrium, see Kawai (1983).) The market clearing conditions yield:

$$Z_0 = Q_{0,1} + K_0 - C_{0,1} = C_{1,1} + K_1 - Q_{1,1}.$$

The arguments put forward by Keynes and Hicks assert that production is mostly planned and consumption is largely flexible so that $Q_{0,1} > C_{0,1}$ and $C_{1,1} > Q_{1,1}$. From this, $Z_0 > 0$ follows and there exists a 'normal backwardation' (or a positive risk premium). But when the adjustment cost of changing production is low and that of changing consumption high, such technological conditions may not be satisfied. Furthermore, in some markets (such as those for foreign exchange and financial instruments) the technological distinction between production (Q) and consumption (C) is unimportant and storage can be negative ($K < 0$); then, normal backwardation is not guaranteed. In essence, whether or not normal backwardation is generated depends on the nature of the commodities in question and is an empirical matter.

Considerable empirical effort has been devoted to detecting a positive or negative risk premium in various types of markets, with mixed result (see Peck 1977). The 'efficient futures market hypothesis' (the hypothesis of no systematic risk premium combined with rational expectations) cannot be rejected for many markets, thus invalidating the theory of normal backwardation. In other markets, time-varying risk premia, positive or negative, have also been found.

See Also

- ▶ Futures Trading
- ▶ Spot and Forward Markets

Bibliography

Hicks, J.R. 1946. *Value and capital*, 2nd ed. London: Oxford University Press.
Kawai, M. 1983. Price volatility of storable commodities under rational expectations in spot and futures markets. *International Economic Review* 24(2): 435–454.
Keynes, J.M. 1923. Some aspects of commodity markets. In *The Manchester guardian commercial, reconstruction supplement*, 29 Mar. Reprinted in *The collected writings of John Maynard Keynes*, vol. 7. London: Macmillan.
Keynes, J.M. 1930. *A treatise on money*, vol. 2. London: Macmillan.
Peck, A.E. (ed.). 1977. *Selected writings on futures markets*, vol. 2. Chicago: Chicago Board of Trade.

Backwardness

M. Falkus

The term 'economic backwardness' is frequently used as a synonym for 'economic underdevelopment' and in this sense was first used by John Stuart Mill in the 1850s. Since 1950, however, the concept of 'relative economic backwardness', whereby characteristics of the development process are seen to be in the level or stage of development reached by a particular country, has come to be associated with the ideas put forward by Alexander Gerschenkron. It is Gerschenkron's concept which will be considered here.

The hypothesis that a nation's relative economic backwardness helps shape the contours of its subsequent development has a lengthy history. Versions of such a concept can be found in a number of 19th-century writings, most explicitly in relation to Russia. Thus both Herzen and Chernyshevskii, for example, specifically linked the expected path of Russia's industrialization with her level of backwardness. Although Gerschenkron himself made fullest use of his hypothesis in his writings on Tsarist Russia's industrial development, he never discussed the historical antecedents of his theories. Moreover it was Gerschenkron's contribution, not simply to link backwardness and economic change in one country, but to suggest a hypothesis of relative backwardness whereby the entire sequence of industrializing nations in 19th-century Europe fitted into a distinct pattern according to their level of development at the onset of their industrialization.

Gerschenkron first put forward his ideas in an influential essay published in 1952, 'Economic backwardness in historical perspective'. The concept was later refined and elaborated and was most clearly summarized in his 1962 paper 'The approach to European industrialization: a postscript'.

Gerschenkron's hypothesis relates specifically to the pattern of European industrialization in the 19th century. The concept of 'relative backwardness' depends first and foremost on the proposition that 'in practice, we *can* rank the countries according to their backwardness and even discuss groups of similar degree of backwardness' (Gerschenkron's italics). Once so ranked a number of further propositions appear. The more backward the country 'the more explosive was the great spurt of its industrialisation, if and when it came'. The pattern of industrialization exhibited by the late starter had a number of characteristics. These characteristics often showed 'the advantages of backwardness', a notion frequently stressed by Gerschenkron. Thus Gerschenkron suggested that the industrial upsurge of the backward late-developer was often associated with modern large-scale plant and enterprise and a tendency among the enterprises to form 'monopolistic compacts' such as trusts and cartels. Capital goods, rather than consumer goods, would dominate the industrial spurt of the late industrializer. The level of backwardness at the onset of industrialization tended to be associated with 'organized direction' of industrial development: the most backward were dominated by state activity, while the moderately backward had their industries largely controlled by investment banks. Also, the more backward the country the less likely was the agricultural sector to play a positive role in the industrialization spurt. Indeed, the industrial spurt would put increasing strains on consumption levels the lower the base from which industrialization started. Gerschenkron also suggested that

as backward countries industrialized, and so became less backward, their patterns of further industrialization took on the character of the less backward: initial diversity gave way to subsequent convergence.

An influential notion inherent in Gerschenkron's hypothesis has been his concept of 'substitutes'; that is, the very backwardness of a country makes it necessary for that country to find substitutes for the internal demand, productive factors, or institutions which the backward country lacks. Thus in Russia the state was a 'substitute' for the entrepreneurial and financial facilities found in the less backward areas. Through the process of substitution, and by developing later, the less developed country could benefit from the 'advantages of backwardness', such as the adoption of the most advanced branches of industry with the latest technology.

Throughout Gerschenkron's writings Tsarist Russia stands as the prime example of the late industrializing backward country, while at the other extreme England was the relatively advanced early industrializer. Thus England's Industrial Revolution was characterized by a slow 'spurt', a concentration on small-scale competitive consumer goods industries, and with individual enterprises rather than banks or the state providing the bulk of industrial finance. Between the two extremes come the 'moderately backward' countries France and Germany, where the activities of investment banks play a crucial role in industrialization and where heavy industries played a larger role in the industrial spurt than they did in England.

The significance of Gerschenkron's scheme rests on several factors. Perhaps most fundamentally Gerschenkron opened up new avenues of historical enquiry by establishing a framework of analysis of differences rather than similarities in the process of modernization. In this way his concept of historical change differed both from Marxian and from other 'stage' theories such as Rostow's. By concentrating on sectoral industrial change rather than on aggregate national accounts Gerschenkron emphasized variables for which more data are available, and this has encouraged the application of statistical techniques to the study of economic history.

As mentioned already, Gerschenkron's major discussions and utilization of the concept of relative backwardness appear in his writings on Russia. He had also applied the concept to a number of individual case-studies which demonstrate how useful the hypothesis may be, even when historical reality does not conform to prior expectations. Of particular note are his studies of industrialization in Italy, Bulgaria, and Austria. For Italy Gerschenkron argued that the rate of industrial growth during the spurt after 1896 was less than might have been anticipated from the initial level of backwardness. The slower pace was due in part to inadequate support from investment banks and the state, and in part because the main burst of railway construction (which might have given impetus to the spurt) had occurred earlier. In Bulgaria, too, the state failed to provide an effective substitute for the lack of internal demand, entrepreneurship, and financial institutions. Austria, argued Gerschenkron, had indeed the potential for a successful industrial spurt in the early years of the 20th century. The problem here was that, in contrast to Witte's Russia, the state was divided against itself: the Ministry of Finance obstructed the efforts of Prime Minister Koerber to introduce schemes for promoting large-scale industry, and, as in Italy, there had already been some measure of modern industrialization before the spurt.

These case studies are useful for the insights they provide into the process of industrialization even where the pattern suggested by Gerschenkron's scheme fails to materialize fully. Another brilliant application of the concept of relative backwardness was provided in Gerschenkron's (1970) study of European mercantilism, showing how the most backward countries around the beginning of the 18th century were those where mercantilist policies were most fully developed and applied.

Gerschenkron's hypothesis has come under critical scrutiny from several quarters. Some suggest that the concept of relative backwardness

itself is too general and vague to be measured and tested in a meaningful way, although Sandberg has endeavoured to refine the concept by separating those countries backward through 'poverty' and those backward through 'ignorance' (a low level of educational attainment). Barsby has pointed out that several of Gerschenkron's key suggestions, such as the greater role of the state and of modern large-scale heavy industry in the industrial spurts of backward countries, are empirically difficult to determine; while Good has shown that the role of banking in European industrialization does not always conform to the Gerschenkronian pattern. Challenges, too, have come from historians of Russia. The roles played by the state, banks and agriculture in Russian industrialization suggested by Gerschenkron have been called into question (by Gregory and Crisp among others), while the suggested convergence of the Russian pattern towards that exhibited by less backward nations has also been denied. It has been noted, too, that Gerschenkron's hypothesis makes the nation-state rather than the region the unit of economic analysis, while other critics argue that Gerschenkron ignores such influences as military expenditure and the particular conjunction of railway and iron and steel development which influenced European growth rates at the close of the 19th century.

Literature specifically concerned with Gerschenkron's hypothesis, however, whether critical or otherwise, is an inadequate guide to the influence of the concept of relative backwardness on historiography. Gerschenkron's approach has proved both fruitful and enduring. Following Rosovsky's pioneering attempt to apply the concept to Japan, a number of studies have used the hypothesis of relative backwardness to analyse growth patterns in Africa, Asia, and elsewhere. Evidently, the approach will have widespread application far beyond the temporal and geographical limits set by Gerschenkron himself. Indeed, Gerschenkron's outstanding intellectual legacy may well lie not so much in his own studies of 19th-century European industrialization, which are increasingly subject to criticism and reinterpretation, but in the development of a major heuristic framework which will continue to provide insights into patterns of economic development across a wide spectrum of societies and time periods.

See Also

- ▶ Catching-Up
- ▶ Cumulative Causation
- ▶ Development Economics
- ▶ Gerschenkron, Alexander (1904–1978)
- ▶ Industrial Revolution
- ▶ Periphery

Bibliography

Barsby, S.L. 1969. Economic backwardness and the characteristics of development. *Journal of Economic History* 29(3): 449–472.

Cameron, R. 1972. *Banking and economic development: Some lessons of history*. New York: Oxford University Press.

Crisp, O. 1976. *Studies in the Russian economy before 1914*. London: Macmillan.

Gerschenkron, A. 1952. Economic backwardness in historical perspective. In Gerschenkron (1962a).

Gerschenkron, A. 1955. Notes on the rate of industrial growth in Italy, 1881–1913. In Gerschenkron (1962a).

Gerschenkron, A. 1962a. *Economic backwardness in historical perspective*. Cambridge, MA: Harvard University Press.

Gerschenkron, A. 1962b. Problems and patterns and Russian economic development. In Gerschenkron (1962a).

Gerschenkron, A. 1962c. Some aspects of industrialization in Bulgaria, 1878–1939. In Gerschenkron (1962a).

Gerschenkron, A. 1967. The discipline and I. *Journal of Economic History* 27(4): 443–459.

Gerschenkron, A. 1968. *Continuity in history and other essays*. Cambridge, MA: Harvard University Press.

Gerschenkron, A. 1970. *Europe in the Russian mirror: Four lectures in economic history*. Cambridge: Cambridge University Press.

Gerschenkron, A. 1977. *An economic spurt that failed: Four lectures in Austrian history*. Princeton: Princeton University Press.

Good, D.F. 1973. Backwardness and the role of banking in nineteenth century European industrialisation. *Journal of Economic History* 33(4): 845–850.

Gregory, P. 1973–1974. Some empirical comments on the theory of relative backwardness: The Russian case. *Economic Development and Cultural Change* 22(4): 654–665.

Rosovsky, H. 1961. *Capital formation in Japan, 1868–1940*. New York: The Free Press.

Sandberg, L. 1982. Ignorance, poverty and economic backwardness in the early stages of European industrialisation: Variations on Alexander Gerschenkron's grand theme. *Journal of European Economic History* 11(3): 675–698.

Bagehot, Walter (1826–1877)

Asa Briggs

Keywords

Bagehot, W.; Bank of England; Bimetallism; British classical economics; Expectations; Free trade; Giffen, R.; Jevons, W. S.; Mathematics and economics; Mill, J. S.; Ricardo, D.; Socialism; Statistics and economics; Trade unions; Walras, L

JEL Classifications

B31

Editor and literary critic as well as banker and economist, Bagehot was described in retrospect by Lord Bryce as 'the most original mind of his generation' (Buchan 1959, p. 260). It is a difficult claim to sustain, certainly as far as his scattered economic writings are concerned. There was no doubt, however, about his intellectual versatility: there was an immediacy, a clarity and an irony – what he said of his friend Arthur Hugh Clough's poems, 'a sort of truthful scepticism' – about Bagehot's essays in different fields which make them still pre-eminently readable. Bagehot saw connections, too, between economics, politics, psychology, anthropology and the natural sciences – 'mind and character' – refusing to draw rigid boundaries between most of these subjects and 'literary studies', while recognizing in his later years that the frontiers of political economy needed to be more carefully marked. 'Most original' or not, he was, as the historian G.M. Young (1948) has observed, *Victoranum maxime,* if not *Victoranum maximus*: 'he was in and of his age, and could have been of no other.' He pre-dated academic specialization and professionalization, and he was never didactic in his approach.

His first writing on economics, a revealing if not a searching review of John Stuart Mill's *Principles of Political Economy,* appeared in 1848 before the sense of a Victorian age had taken shape. His last and most voluminous writing on the subject appeared posthumously in a volume of essays, the first on 'the postulates of English political economy', which his editor-friend Richard Holt Hutton entitled *Economic Studies* (1879). By then the economic confidence of the mid-Victorian years was over, and there were many signs both of economic and social strain, some of which Bagehot had predicted. It was in 1859, the *annus mirabilis* of mid-Victorian England, however, the year of Darwin's *Origin of Species,* Mill's *On Liberty* and Smile's *Self Help,* that Bagehot became editor of *The Economist,* a periodical founded by his father-in-law James Wilson, and it was through his lively editorship, which continued until his death, that he was in regular touch with an interesting and influential, if limited, section of his contemporaries. 'The politics of the paper', he wrote simply, 'must be viewed mainly with reference to the tastes of men of business.'

The mid-Victorian years constituted, in his own phrase, 'a period singularly remarkable for its material progress, and almost marvellous in its banking development'. It was the latter aspect of the period which provided him with the theme of his best-known and brilliantly written book *Lombard Street,* which was begun in 1870 and appeared in 1873. It dealt, however, as it was bound to do, not only with the 'marvellous development', but with the 'panics' of 1857 and 1866 to which the Bank of England, the central institution in the system, had to respond. Indeed, the germ of *Lombard Street* was an article written in *The Economist* in 1857, 13 years after Peel's Bank Charter Act, and it was in 1866 that he took up the theme again.

Bagehot's conviction that the Bank of England neither fully understood nor fully lived up to its responsibilities was the product of years of experience which went back to his own early life between 1852 and 1859 as a country banker with Stuckey's at Langport, his birthplace, in the West of England, where his father also was a banker. The chapter on deposit banking reflects this. So, too, does his complaint that the directors of the Bank of England were 'amateurs', and his insistence that the 'trained banking element' needed to be augmented.

Lombard Street is a book with a distinctive purpose rather than an essay in applied economics; and, as Schumpeter has observed, 'it does not contain anything that should have been new to any student of economics'. The main stress in it is on confidence as a necessary foundation of London's banking system. 'Credit – the disposition of one man to trust another – is singularly varying. In England after a great calamity, everybody is suspicious of everybody; as soon as that calamity is forgotten everybody again confides in everybody.' Bagehot underestimated the extent to which through joint stock banks' cheques trade was expanding without increases in note issue and the extent to which the Bank of England itself was beginning to develop techniques of influencing interest rates. He also overestimated the extent to which in 'rapidly growing districts' of the country 'almost any amount of money can be well employed'. In the last resort, too, his policy recommendations were deliberately restricted. He was disposed in principle to a 'natural system' in which each bank kept its own reserves of gold and legal tender, but in English circumstances he saw no more future in seeking to change the system fundamentally than in changing the political system. 'I propose to retain this system because I am quite sure that it is of no manner of use proposing to alter it.' With a characteristic glance across the Channel to France for a necessary comparison – things were done very differently there – he noted how the English system had 'slowly grown up' because it had 'suited itself to the course of business' and 'forced itself on the habits of men'. It would not be altered, therefore, 'because theorists disapprove of it, or because books are written against it'.

Bagehot had little use for 'theorists' and disdained the French for what he called their 'morbid appetite for exhaustive and original theories'. He described political economy 'as we have it in England' as 'the science of business' and did not object to the fact that it was 'insular'. Yet he talked of the 'laws of wealth' and believed that they had been arrived at in the same way as the 'laws of motion'. Free trade was such a law. It was impossible, he argued, to write the history of ' similar phenomena like those of Lombard Street' without 'a considerable accumulation of applicable doctrine': to do so would be like 'trying to explain the bursting of a boiler without knowing the theory of steam', a not very helpful analogy since the invention of the steam engine preceded the discovery of the laws of thermodynamics. Bagehot relied considerably on analogies. 'Panics', for example, were 'a species of neuralgia'. The 'unconscious "organization of capital"' in the City of London, described by Bagehot as a 'continental phrase', depended on the entry into City business of a 'dirty crowd of little men'; and this 'rough and vulgar structure of English commerce' was 'the secret of its life' because it contained 'the propensity to variation' which was 'the principle of progress' in the 'social as in the animal kingdom'.

Such an approach to political economy was radically different from that of W.S. Jevons who, like Bagehot, had been educated at University College, London, or 'M. Walras, of Lausanne' who, according to Bagehot himself, had worked out 'a mathematical theory' of political economy 'without communication and almost simultaneously'. There were however three defects, Bagehot maintained, in the British tradition of political economy, which started with Adam Smith but was sharpened and 'mapped' by David Ricardo. First, it was too culture-bound; for example, it took for granted the free circulation of labour, unknown in India. Second, its expositors did not always make it clear that they were dealing not with real men but with 'imaginary' ones. Abstract political economy did not focus on 'the entire man as we know him in fact, but . . . a man answering to pure definition from which all impairing and conflicting elements have been fined away'. It was not concerned with 'middle

principles'. Third, considered as a body of knowledge, English political economy was 'not a questionable thing of unlimited extent but a most certain and useful thing of limited extent'. It was certainly not 'the highest study of the mind'. There were others 'which are much higher'.

Bagehot did not push such criticism far. He had much to say about primitive and pre-commercial economies, but he put forward no theory of economic development. Nor, despite an interest in methodology, did he draw out the full implications of his own behaviourist (and in places institutionalist) approach to economics. Finally, he offered no agenda for political economists in the future. He noted, as others noted, that during the 1870s political economy lay 'rather dead in the public mind. Not only does it not excite the same interest as it did formerly, but there is not exactly the same confidence in it.' His own preoccupations in that decade were more practical than theoretical despite the writing of such essays as 'The Postulates of English Political Economy', which first appeared in article form in the *Fortnightly* in 1876. He never completed a new essay on Mill, and an essay on Malthus, whom he took along with Smith, Ricardo and Mill to be the founders of British political economy, revealed more interest in the man than in his thought. In the year when the 'Postulates' appeared, he successfully suggested to the Chancellor of the Exchequer the value to the Treasury of short-term securities resembling as much as possible commercial bills of exchange. The result was the Treasury Bill. The fact that the Chancellor was then a Conservative mattered little to the liberal-conservative Bagehot, who was described by his Liberal admirer W.E. Gladstone as a 'sort of supplementary Chancellor of the Exchequer'.

Bagehot was as out of sympathy with the liberal radicals of the 1870s as he was with the bimetallists, and he had never shown any sympathy for socialist political economy. He saw the capitalist as 'the motive power in modern production' in the 'great commerce', the man who settled 'what goods shall be made, and what not'. Nonetheless, he stated explicitly in several places that he had 'no objection whatever to the aspiration of the workmen for more wages', and he came to appreciate more willingly than Jevons the role of trade unions and collective bargaining. In his first review of Mill in 1848 he had stated that 'the great problem for European and especially for English statesmen in the nineteenth century is how shall the [wage] rate be raised and how shall the lower orders be improved'. Some of the views he expressed on this subject – and on expectations – were not dissimilar to those of the neoclassical Alfred Marshall. He did not use the term 'classical' himself in charting the evolution of British political economy.

Bagehot left no school of disciples. He was content to persuade his contemporaries. His sinuous prose style was supremely persuasive. So, too, was his skill in sifting and assessing inside economic intelligence. Yet while he devoted little attention to precise quantitative evidence in *Lombard Street* and, unlike Jevons, saw little point in developing economics in mathematical form, he was always interested in numbers as well as in words. One of his closest collaborators on the staff of *The Economist*, the statistician Robert Giffen, his first full-time assistant, paid tribute to 'his knowledge and feeling of the "how much" in dealing with the complex workings of economic tendencies'. 'He knew what tables could be made to say, and the value of simplicity in their construction.' Bagehot always maintained, however, that while 'theorists take a table of prices as facts settled by unalterable laws, a stockbroker will tell you such prices can be made'. Statistics were 'useful': they needed to be interpreted by 'men of business' who possessed the grasp of 'probabilities' and the 'solid judgement' which Bagehot most admired and which he sought to express. Indeed, business for him was 'really a profession often requiring for its practice quite as much knowledge, and quite as much skill, as law and medicine'. Businessmen did not go to political economy: political economy, as in the case of Ricardo, came to them.

Selected Works

All Bagehot's economic writings are collected. In *The collected works of Walter Bagehot*, ed. N. St. John Stevas, vols. 1–15 (1978–86). London: The Economist.

Bibliography

Buchan, A. 1959. *The spare chancellor. The life of Walter Bagehot*. London: Chatto & Windus.
Giffen, R. 1880. Bagehot as an economist. *The Fortnightly,* April, 549–567.
Young, G.M. 1948. The greatest Victorian. In *Today and yesterday*, ed. G.M. Young. London: Rupert Hart-Davis.

Bailey, Samuel (1791–1870)

R. M. Rauner

Keywords

Absolute and exchangeable value; Cotterill, C. F.; De Quincey, T.; Index numbers; Labour supply; Labour theory of value; Lowe, J.; McCulloch, J. R.; Macleod, H.D.; Malthus, T. R.; Mill, J.; Mill, J. S.; Relative value; Ricardo, D.; Scrope, P.; Seligman, E. R. A.; Torrens, R

JEL Classifications

B31

Samuel Bailey was born in Sheffield, England, one of 11 children. His father was a cutler and merchant of substance. Samuel also became a merchant and banker. Throughout his life, he served on the Sheffield Town Trust (a quasi-governmental agency) and was twice a candidate for Parliament in the Reform elections of 1832 and 1835. Writing widely on banking, politics and philosophy, he lived his entire life in Sheffield, unmarried, and died there in 1870.

Bailey published his principal economic work, *A Critical Dissertation*. . ., in 1825, a time when Ricardian theory was nearing its peak of popularity and acceptance. The *Westminster Review* (1826) thought the *Critical Dissertation* inconsequential, and J.R. McCulloch (1845) later claimed that it had not shaken the foundations of Ricardo's labour theory of value. Robert Torrens, however, praised Bailey's book in 1831 at the London Political Economy Club, and John Stuart Mill brought it before his bi-weekly reading group. This attention, nevertheless, did not keep Bailey on front stage, and he had to be rediscovered later by E.R.A. Seligman (1903); the London School of Economics republished the *Critical Dissertation* in 1931. Schumpeter (1954) judged Bailey's tract to be a 'masterpiece of criticism' and to lie near the 'front rank in the history of scientific economics'. R.M. Rauner (1961) re-examined Bailey's work from a larger perspective.

The centrepiece of Bailey's argument was his definition of value as ultimately 'esteem' or a 'mental affection'. The 'specific feeling of value', however, arose only when items were subject to preference or exchange. This defined value as relative, not something intrinsic like labour in Ricardo's theory. Value is the amount of one commodity exchanged for another; it is measured in terms of a third commodity with which the two exchange if they are not directly bartered. From this position Bailey attacked Ricardo's postulate that labour effort defined value. He showed that, despite Ricardo's claim to the contrary, constancy of labour used in production could not assure constancy in exchange value – unless value were defined differently. This, of course, is what Ricardo had done in shifting from exchange to 'real' or 'absolute' value.

Ricardo's conception of value as an absolute and his endless search for a standard of invariable value opened him to Bailey's stricture that constancy of value meant constancy in exchange ratios. Evidence and observation showed that exchange values rarely stayed constant. To the Ricardians, however, constancy of value meant constancy of labour cost of production; this, they believed, was necessary in the determination of whether individual economic welfare had changed over time. Bailey objected that exchange of commodities cannot take place between two different time periods. Exchanges occur at different times and these exchanges can be compared. But such comparisons are the only way economic welfare in different times or places can be assessed. In a later tract (1844), Bailey used this

same argument, making the point that interperiod contracts could be fixed only in terms of quantities, not constancy of values. This enabled him to oppose the index number proposals (then called 'tabular standards') of Joseph Lowe and Poulett Scrope. Such standards could not assure constancy of quantities exchanged in different times, a criticism of index numbers that is still valid today.

Using relative value as his anchor, Bailey then demonstrated that Ricardo's theory of wages was faulty. He insisted that labour value – wages – was definitionally the same as all other value, namely, what a unit of labour exchanged for. Ricardo's theorem, that wages and profits varied inversely, was wrong since it implied that wages could be high (i.e. taking a large proportionate share of production) while labour value was low, wages exchanged for little and workers were near starvation.

The relative value concept applied to wages allowed Bailey an easy application of the principles of rent to labour. Just as with land, different values for labour were caused by the monopoly characteristics of labour supply, as well as by differential productivity due to varying labour skill or dexterity. This contrasted sharply with Ricardian–Malthusian subsistence wages. Unfortunately, Bailey did not use the same reasoning against capital and merely denoted profits as the gain over capital employed.

The *Critical Dissertation* prompted some serious attempts to clear up the loose ends in Ricardo, most notably by McCulloch (1845); by the anonymous *Westminster Review* article (1826), probably written by James Mill (1826); and by Thomas De Quincey (1844). But Ricardo's *system* held fast. Malthus (1827) devoted the largest part of his work on definitions to Bailey, mainly quarrelling over the purely relative value notion. He reaffirmed the importance of a constant, unvarying measure of value, defined as the quantity of labour commanded by commodities in exchange. Samuel Read (1829) drew on Bailey's destruction of the Mill-McCulloch theory that time used in production is congealed labour, but he did not follow Bailey on the relativity of value or the measure of value. C.F. Cotterill (1831) and H.D. Macleod (1863, 1866) both praised Bailey's work and used his treatment of the nature and measure of value in their own studies.

From a larger perspective, by stressing relative value exclusively, Bailey pulled economic analysis back from the Smith-Ricardo stream that sought a principal cause of value to explain the production and distribution of material wealth among the labouring, rentier and capitalist classes. In Bailey's argument relative values – prices – vary for all kinds of reasons affecting demand ('esteem') and supply (production under constant or increasing cost, supply-limiting) conditions. Hence, his view involves no notion of long-run growth, tendencies toward equilibrium, stationary states or other systemic visions. Everything is relative; individual economic welfare is expressed period-by-period solely in terms of relative values.

Bailey's is an incomplete treatment if one demands that value theory be integral with the determination of social, institutional and economic forces in an interdependent production system. On the other hand, Bailey's work freed analysis from the need to link production and distribution to socioeconomic class relationships. It pointed instead towards relationships between individual needs and perceptions, and the material goods that can satisfy them.

Selected Works

1821. *Essays on the formation and publication of opinions and other subjects*. London.
1823. *Questions on political economy, politics, metaphysics, polite literature and other branches of knowledge*. London.
1825. *A critical dissertation on the nature, measures and causes of value; chiefly in reference to the writings of Mr. Ricardo and his followers*. London.
1826. *A letter to a political economist; occasioned by an article in the Westminster review on the subject of value*. London.
1830. *A discussion of parliamentary reform*. London.
1835. *The rationale of political representation*. London.

1837. *Money and its vicissitudes in value; as they affect national industry and pecuniary contracts; with a postscript on joint-stock banks*. London.

Bibliography

Anon. 1826. Letter to a political economist. *Westminister Review,* January.
Cotterill, C.F. 1831. *An examination of the doctrines of value as set forth by Adam Smith, Ricardo, McCulloch, the author of 'A Critical Dissertation', etc., Torrens, Malthus, say, etc., being a reply to those distinguished authors*. London.
De Quincey, T. 1844. *The logic of political economy*. Edinburgh: William Blackwood and Sons.
Macleod, H.D. 1863. *A dictionary of political economy*. London.
Macleod, H.D. 1866. *The theory and practice of banking*. 2nd ed. London: Longmans, Green, Reader, & Dyer.
Malthus, T.R. 1827. *Definitions in political economy*. London: Murray.
McCulloch, J.R. 1825. *Principles of political economy*. London.
McCulloch, J.R. 1845. *The literature of political economy*. London: Longman, Brown, Green & Longmans.
Mill, J. 1826. *Elements of political economy*. 3rd ed. London: Baldwin, Cradock, and Joy.
Rauner, R.M. 1961. *Samuel Bailey and the classical theory of value*. London: London School of Economics and Political Science, G. Bell & Sons.
Read, S. 1829. *An inquiry into the natural grounds of right to vendible property or wealth*. Edinburgh: Oliver and Boyd.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Seligman, E.R.A. 1903. On some neglected British economists. *Economic Journal* 13 (335–63): 511–535.

Bain, Joe Staten (1912–1991)

William G. Shepherd

Keywords

Bain, J. S.; Barriers to entry; Concentration; Economies of scale; Industrial organization; Limit pricing; Market power; Market share; Oligopoly group; Profitability

JEL Classifications

B31

Joe S. Bain was born in Spokane, Washington, on 4 July 1912. After graduating from the University of California at Los Angeles in 1935 and gaining the doctorate from Harvard in 1940 (under Joseph Schumpeter), he spent his entire career at the University of California at Berkeley, retiring in 1975. He was appointed Distinguished Fellow of the American Economic Association in 1982.

A prolific and seminal writer, Bain helped to shape the field of industrial organization in its modern form, with special attention to market structure. Bain's analysis focused on the oligopoly group within an industry, and on barriers to new competition. He also worked on natural resource development by public enterprise, concentrating on the oil industry.

Bain's empirical work on economies of scale, entry barriers, and limit pricing broke new ground. He developed the field's intellectual format, in which technical factors may determine structure, and structure then influences behaviour and performance. Some of these concepts were already current as early as 1900. During 1925–40, as the field took shape, attention shifted to the industry and the oligopoly group within it.

In the 1930s, Bain entered a formative field which was rich in possibilities for giving new rigour to older concepts, for developing new ones, and for shaping the framework. That has been his main role and contribution. Though he did not create concepts, nor indeed the framework, he selected from among them and carried their scientific analysis further than anyone else.

The analysis grew after 1940 in a series of articles and chapters, culminating in *Barriers to New Competition* in 1956 and *Industrial Organization* in 1959. His analysis was verbal and graphical rather than mathematical. In Bain's analysis of the conditions of entry, the barriers have three possible economic sources; absolute cost advantages, product differentiation and size. Barriers then permit 'limit pricing' by a firm or firms which consciously apply their strategy towards entry.

Bain drew the main conclusions, and he noted the difficulties of empirical tests. The definition of barriers as a single, general phenomenon posed special problems, which are still unsolved. Since 1960, over seven new barrier 'sources' have been proposed, and the concept of barriers has tended to acquire just that ad hoc character which Bain frequently reproved in others' theories.

Measurement has also proven to be difficult. It requires a merging of disparate objective and subjective data about the barriers' causes. Whether these sources of barriers are additive, multiplicative or merely parallel was also left unclear by Bain (and all others).

Bain's measurement of scale economics was pioneering. Earlier studies had suffered from data problems and from a mingling of technical and pecuniary elements. Bain centred unerringly on technical economies. Thereby he gave the first solid normative basis for evaluating excess concentration.

By estimating 'best practice' conditions for scale for new capacity, Bain neatly avoided the normative–positive confusion which infects cross-section studies of past costs and survivor tests of emerging sizes. His 'engineering' estimates supply a normative basis for appraising how much concentration is socially 'necessary'.

Profitability was also analysed closely by Bain. He tried nearly every available method to factor out the concentration–profitability relationship. In a 1949 article (later extended in *Barriers*), Bain put the study of profitability on a firm scientific and normative basis. His findings of a step function, with a break at 70 per cent for eight-firm concentration, has tended to be replaced in recent research by a continuously sloping concentration–profitability relationship. Still, Bain set the basis for all good later research on the subject.

Bain's architectural choices in using and emphasizing individual elements were distinctive. Three features stand out – the triad, the industry basis, and the stress on the oligopoly group behind an entry barrier. (1) Bain developed the three-tier format of structure, behaviour and performance with what may be called a 'soft structuralist' emphasis. Bain used it as a broad set of concepts, by which the whole subject (theory, tests, policy lessons) is organized, not as just a format for individual cases. (2) Bain used the *industry* as the basic unit behaviour. It was a choice that shaped the images and methodology in distinctive ways. (3) The *oligopoly group*, setting limit price strategy behind an entry barrier, came to be the most distinctive part of Bain's analysis. As of 1949–50, Bain regarded concentration as the key determinant of market power and profitability.

By 1951 he appeared to regard barriers as the decisive element, which could be both necessary and sufficient to govern profitability. Yet Bain later suggested frequently that barriers would be highly correlated with the degree of concentration. In fact, all of the sources of barriers are also sources of high market shares and concentration. Do barriers shape the dominant firm's share, or do they operate jointly?

Any eventual resolution of barriers' role will probably assign barriers at least a significant role, thanks to Bain's stress on them. He put the concepts and relationships in testable form, and he began the testing of them. To a large extent he rescued the subject from a preoccupation with oligopoly interactions and games, and he gave it a strong framework.

Yet Bain's most durable contribution lies deeper, in the methods and research standards of the field. By 1960, he had helped to give it structure, precision, and high standards of research quality. He selected the main concepts and relationships, gave them extended analysis, tested them, and drew policy lessons. The individual parts were related within a framework of causation and performance.

His more specific methods and results have also continued to be valid because they met these standards. Beyond the individual concepts and tests is the fact that they fit together in a system, and that this system was carefully developed and tested. That is the way to scientific permanence.

Selected Works

1944, 1945, 1947. *The economics of the pacific coast petroleum industry*. 3 vols. Berkeley: University of California Press.

1948. *Pricing, distribution and employment: Economics of an enterprise system*. New York: Henry Holt; Italian and Spanish translations.
1956. *Barriers to new competition: Their character and consequences in manufacturing industries*. Cambridge, MA: Harvard University Press.
1959. *Industrial organization*. New York: Wiley; revised edn, 1968.
1966a. *International differences in industrial structure: Eight nations in the 1950s*. New Haven: Yale University Press.
1966b. (With R.E. Caves and J. Margolis.) *Northern California's water industry: The comparative efficiency of public enterprise in developing a scarce natural resource*. Baltimore: Johns Hopkins Press.
1972. *Essays on price theory and industrial organization*. Boston: Little, Brown & Co.

Bairoch, Paul (1930–1999)

Elise S. Brezis

Keywords

Agricultural revolution; Bairoich, P; City and economic development; Cliometrics; Colonialism; Diffusion of technology; Economic development; Foreign aid; Free trade; Inequality (global); Protection; Technological progress; Trade and economic growth; Urbanization

JEL Classifications

B31

Paul Bairoch was born in Antwerp in 1930. He was the son of a Jewish family that emigrated from Poland to Belgium in the 1920s, and that later went into exile in a small village in the Gers, France, during the Second World War. After the war, Bairoch moved to Brussels, later spent a short period in Israel, and upon his return to Belgium began to study economic history. While a research fellow at the University of Brussels, Bairoch developed statistical time series on the national statistics of Belgium, worked on his doctorate, and in 1963, presented his thesis, 'The Starting Process of Economic Growth'. He then went on to teach in a number of universities and even worked at General Agreement on Tariffs and Trade (GATT) for a time. From 1972 onwards, Bairoch was a member of the faculty at the University of Geneva, where he was director of the Center of International Economic History until his death in 1999.

A trait common to all Bairoch's research in economic history from his thesis onwards was that he based his opinions on data, and, when the data did not exist, he found a way to collect or construct new data. Bairoch can be seen as a pioneer of cliometrics, and believed that economic history cannot survive without data and statistical information. David Landes (1998, p. xiii) even gave Bairoch the nickname 'collector and calculator of the numbers of growth and productivity'. Another characteristic typical of Bairoch's research is that he was not afraid to be nonconformist and present views that ran against the mainstream.

Bairoch worked in three main subjects: economic development and growth, urban studies and international trade.

Population, Cities, and Urban Research

Bairoch was interested in the relationship between urbanization and economic development, and examined urban evolution from the Neolithic period to 1900. He developed series on sizes of cities from AD 800 to 1850.

Bairoch's main achievement in this field was showing that there was a typical pattern of urbanization: traditional societies reached their maximum urban population rapidly, levelling off at somewhere between 8 and 15 per cent (Europe reached this level around 1300), and maintained this proportion until the onset of industrialization,

when the urban population then surged. He also observed that for non-developed countries urbanization has negative consequences for agricultural development.

Development, Industrialization, and Inequality

One of the main topics of Bairoch's research was the dynamics of development and the inequality between developed and developing countries. In his last book, *Victoires et déboires* (1997), a formidable synthesis of the economic and social history of the world, Bairoch tried to explain the pre-eminence of the West, and the setbacks *(déboires)* suffered by the Third World.

Regarding the mechanism of development of the West, Bairoch insisted on the necessity of an agricultural revolution, and also on the importance of institutions. He had also a strong interest in the development of technological progress in the 19th century, and stressed the differences between it and the diffusion of the science-based technology of the 20th century.

Bairoch also analysed at length the reasons for the backwardness of the Third World, and through the use of comparative statistics his analysis includes a comparison between its present economic progress and that of developed countries at the times of their take-offs. Bairoch's conclusions were that the absence of an agricultural revolution and failure to reduce fertility rates were among the most binding facts impeding development. He was therefore pessimistic about the prospects for development of the lagging countries, especially those in Africa.

Regarding inequality, Bairoch stressed that before the Industrial Revolution no appreciable difference in per capita income separated western Europe from the rest of the world, while the gap between the developed and the developing world increased thereafter. Moreover, regarding the effect of colonialism, Bairoch stressed that colonialism was not only largely unprofitable for the West but also harmed the Third World. Bairoch was a proponent of foreign aid to reduce inequalities.

International Trade

Probably Bairoch's best-known work is *Economics and World History: Myths and Paradoxes* (1993), in which he sets the record straight on 20 commonly held myths about economic history, among them that free trade has historically led to periods of economic growth; a myth associated with those who 'could be described as a conservative group that romanticizes the 19th century and makes free trade almost into a sacred doctrine' (1993, p. xiv).

Bairoch claimed that the idea that free trade was the rule during the 19th century is a myth based on insufficient knowledge and misguided interpretations of the economic history of the United States, Europe, and the Third World, since protection is the rule and free trade the exception. Moreover, Bairoch expressed doubts that free trade leads to economic growth. His thesis was that during development countries use protectionist policies, which they dismantle once they industrialize. He showed that Britain protected its home market until British firms in the main sectors dominated the market, and only later on did Britain advocate free trade.

I cannot conclude without mentioning Bairoch's personality: he combined the best of open-minded curiosity and a powerful intellect with warmth, humanity and overwhelming kindness to all who knew him.

See Also

- ▶ Development Economics
- ▶ Economic History
- ▶ International Trade Theory

Selected Works

1975. *The economic development of the Third World since 1900.* London: Methuen.
1988. *Cities and economic development from the dawn of history to the present.* Chicago: University of Chicago Press.

1993. *Economics and world history: Myths and paradoxes.* Chicago: University of Chicago Press.
1997. *Victoires et déboires: Histoire economique et sociale du monde du 16ème siècle à nos jours,* 3 vols. Paris: Gallimard.

Bakunin, Mikhael Alexandrovitch (1814–1876)

David Clark

Mikhail Alexandrovitch Bakunin was unique amongst 19th-century revolutionaries. He combined a deep interest in political theory, philosophy and political economy with a love of political action. Not satisfied to merely outline the evils of existing society and draw blue-prints of superior ones, he propagandized, formed political secret societies and supported every political upheaval, large or small, hopeful or doomed to failure, of his era. Alexander Herzen said of him, 'Everything about this man is colossal, his energy, his appetite, yes, even the man himself.'

Bakunin was born in the Novotorschok district of Tver province. From his father he inherited an intellectual interest in the Encyclopédists and the ideas of Jean-Jacques Rousseau; from his pious sisters an interest in the cult of the inner life. At the age of 14 he was sent to the artillery school in St Petersburg but resiled against a military career and began studying Goethe, Schiller and Fichte – and later Hegel. Along with many other Russians of his generation, he was also deeply influenced by his fellow-Russian, Vissarion Belinsky, who preached a love of the poor.

After a period in Berlin, Dresden and various Swiss cities he moved to Paris in 1844, where he became acquainted with the French Socialists and various Russian exiles. At this time he was particularly influenced by Proudhon. As a result of Bakunin's calls for a revolution in his homeland and the establishment of a republican federation of all Slavic countries, Nicholas I issued a decree depriving him of all his civil, property and nobility rights and sentencing him to lifelong exile in Siberia should he ever return to Russia. Arrested in 1849 during the Dresden Revolution, he was sentenced to death, chained to a wall for a over a year, then returned to St Petersburg where he was kept in solitary confinement. Here he was pressured to write his famous *Confession*, on the insistence of the tsar. In 1857, when he was in bad health and close to suicide, his family succeeded in having him exiled to Siberia, from which he escaped by sailing down the Amur River. Over the next few years, he toured Europe and set up a secret *Fraternité Internationale* of like-minded revolutionaries. His last years were spent in destitution and ill health. Friends reported that his various imprisonments had taken a savage toll on his health. He died in Berne at the age of sixty-two.

Bakunin stood at the cross-roads of several intellectual currents. He was influenced by Slavophilism, Hegelianism, Marxism and Proudhonism. His impact on anarchism was two-tiered. He turned anarchism from a theory of political speculation into a theory of political action. Although he never joined any of the nihilist – 'propaganda by needs' – action groups in Russia or elsewhere, he provided great inspiration for those who did. His writings lacked the ponderous speculation of fellow anarchists such as Godwin, Proudhon or Stirner. They were appeals for action. Hence his most famous 1842 maxim: 'Let us have confidence in the eternal spirit which destroys and annihilates only because it is the unfathomable and eternally creative urge. The urge of destruction is at the same time a creative urge.'

Bakunin also turned anarchism from merely a philosophical position for radical sections of the petty bourgeoisie to a political philosophy which sought mass support from wage earners and the lumpenproletariat, even though its central cadres still tended to come from the intelligentsia. He was a vital influence behind the emergence of organized anarchist movements in Italy, France and Spain in particular in the three decades before World War I.

Bakunin's voluminous writings – Maximoff (1953) provides a good selection – have received

little attention for three main reasons. The first is their fragmentary and issue- and incident-orientated style. The second is his identification with violence. To those unfamiliar with the richness of the anarchist intellectual tradition, Bakunin was far too easily equated with criminals and lunatics. The third reason is his conflict with Marx over the organization and aims of the First International. Marx won this battle but the battle of the giants destroyed the organization. Ever since, Marxist historians and others have done their best to either grossly distort Bakunin's role and position or banish him to historical oblivion, even though his writings deeply influenced both Marx and Lenin. (Bakunin also translated Marx's *Das Kapital* into Russian.)

E.H. Carr's bulky *Michael Bakunin* (1937) devotes litte attention to his ideas but does provide a detailed account of the First International battle. However, the work that is considered Bakunin's best and most mature, *Statism and Anarchism*, is not even mentioned. George H. Sabine's *History of Political Theory* gives him only a passing reference. The best short biography is that by Max Nettlau in Maximoff (1953). Many would consider Sir Isaiah Berlin's assessment of Bakunin – 'He has not bequeathed a single idea worth considering for its own sake; there is not a fresh thought, not even an authentic emotion, only amusing diatribes, high spirits, malicious vignettes, and a memorable epigram or two' (1978, p. 113) – too harsh.

In the history of political theory he remains an enigma. Despite his great emphasis on individual liberty, he believed he had founded a political philosophy to end all political philosophies and demanded unswerving allegiance from his followers. Bakunin will be long remembered for highlighting the greatest political dilemma of our age – how to achieve maximum individual liberty without resort to authoritarian methods and forms of social organization.

See Also

► Anarchism

Selected Works

A complete edition of Bakunin's works has never been published, although a five-volume Russian edition was published by Golas Truda (Moscow and Petrograd, 1919–1922); a three-volume German edition by Verlag der Syndikalist, Berlin, 1921–240; and a six-volume French edition by P.V. Stock, 1895–1913). His *Statism and Anarchism* and *Confessions of a Revolutionary* have been published in many languages. See also A. Lehning (ed.), *Archives Bakounine*, Leiden, 1967.

Bibliography

Berlin, I. 1978. In *Russian thinkers*, ed. H. Hardy and A. Kelly. Harmondsworth: Penguin Books.

Carr, E.H. 1937. *Michael Bakunin*. London: Macmillan.

Maximoff, G.P. (ed.). 1953. *The political philosophy of Bakunin: Scientific anarchism*. London: Collier Macmillan.

Balance of Trade, History of the Theory

S. Bauer

The views of the earliest popular economists of England on the best manner of enriching the nation agree with the measures taken by the legislature and with the balance-of-bargain system, as enforced by the statutes of employment.

> The holl welthe of the reame is for all our riche commodites to gete owt of all other reamys therfore redy money; and after the money is brought in to the holl reame, so shall all peple in the reame be made riche therwith. (Clement Armstrong, *A treatise concerninge the Staple and the Commodities of this Realme*, 1530, pp. 32, 61.)

But when the English merchants had broken down the power of foreign companies and had formed companies of their own, they sought after a rule by which to ascertain what advantages

the regulation of commerce afforded to the nation taken as a whole. Even during the prevalence of the balance-of-bargain system, a rough rule for the policy on which the coinage should be based had been given by an officer of the mint, Richard Aylesbury, who thought that

> provided the merchandise exported from England was properly regulated, that is, if no more of foreign commodities were allowed to be imported than the value of the native commodities which should be taken out, the money in England would remain, and great plenty would come from beyond the seas (*Rolls of Parliament*, vol. iii, p. 126; in Rudding, *Annals of the Coinage*, vol. i, p. 241).

These views, put forward in 1381 by Richard Aylesbury, contrary to the then prevalent opinion (Cunningham, *Growth of English Industry and Commerce, Early and Middle Ages*, 1890, p. 354), were formulated anew and with success by the anonymous author of 'A Discourse of the City of London'. He shows that the increase of prices, which followed the influx of the precious metals from the West Indies had induced the gentry to 'play the fermours, grasiars, brewers, or such like'. This mercantile spirit must be guided by the experience of the merchant's daily practice. England being in need of foreign commodities, and having no mines of its own,

> it followeth necessarily, that if we follow the councel of that good old Husband Marcus Cato, saying, 'oportet patrem familias vendacem esse, non emacem' and do carrie more commodities in value over the seas, then wee bring hether from thence: that then the Realme shall receive that Overplus in Money (*A Discourse of the Names and First Causes of the Institution of Cities, and peopled Towns; and of the Commodities that do grow by the same; and namely of the City of London*, etc. (about 1578), in Stow's *Survey of London*, 1598, p. 450).

William Stafford accepted these principles, adding, that the imported commodities should be 'most apte to be either carried for or kepte in store', and he praised the bailiff of Carmarthen, who had forbidden a ship freighted with oranges to sell them (*A Compendious and Brief Examination*, 1581 edn, New Shakspere Society Edn, pp. 50, 54, 57). This rule of commercial politics has been accepted by John Wheeler (*A Treatise of Commerce*, 1601, pp. 7, 8) and by Gerrard de Malynes, who seems to have suggested the name of balance, saying that the prince should not suffer 'an overbalancing of forreine commodities with this home commodities or in buying more then he selleth' (*A Treatise of the Canker of England's Commonwealth*, 1601, p. 2). The underbalance of trade and the consequent scarcity of money he ascribed to the 'undervaluation of our Money in Exchange', effected by the practices of the bankers. His erroneous ideas and those of Thomas Milles concerning 'merchandising exchange' (*The Customer's Replie*, 1604) were attacked by Edward Misselden, who hoped to remedy this undervaluation of the coin by 'raising' it (*Free Trade or the meanes to make Trade florish*, 1622, pp. 103–5), similar views being expressed in the parliament (*Parliamentary History* vol. i, p. 1195); he calls, however, the balance of trade 'an excellent and politique invention, to shew us the difference of waight in the commerce of one kingdome with another in the scale of commerce' (*The Circle of Commerce, or the Balance of Trade, in defence of Free Trade*, by E.M., 1623, pp. 116, 177). He considers poverty and prodigality as the causes of the present underbalance, the Dutch at once growing rich by manufactures and restraining the home consumption (pp. 132–5). These opinions were generally accepted even by Francis Bacon (*Letter of Advice to George Villiers*, 1616; *Letters and Life*, ed. Spedding, vol. vi, pp. 22–49, and *History of Henry VII*, Works, vol. vi, p. 223), and King James I (*Parliamentary History*, vol. i, p. 1179).

As stress was laid upon the profit of exportation of manufactures, the uselessness of the prohibitions of the exportation of money and bullion became more and more evident. Commercial states like Tuscany and Holland, allowing its free exportation, grew rich, while those forbidding it, like Spain, became impoverished. This point was clearly elucidated by Lewes Roberts, *The Treasure of Traffike*, 1641, p. 77, and the whole doctrine, including the views of exchange as a symptom, not as an agent of trade, as Malynes had maintained, was most systematically explained by Thomas Mun in his posthumous treatise *England's Treasure by Forraigne Trade; or the Balance of our Forraigne Trade is the Rule*

of our Treasure, 1664, who in his *Discourse of Trade* (new edn, 1621) had still advocated the statutes of employment. To him therefore the honour of its invention has often been ascribed. The obstacles to trade were for the most part caused by fiscal motives, and the Commonwealth sought to stimulate the exportation of English commodities by the Act of Navigation. The balance of trade was thought to be advantageous: by fetching the commodities from the immediate places of their production and by sending them to their best market, where they yield the greatest price, but above all by the cheapness of the exported manufactures and the reduction of the price of labour (*The Advocate: or a Narrative of the State and Condition of Things between the English and Dutch Nation*, 1651 edn). This programme was supported by the greatest economists of the end of the 17th century like Petty, Temple, Locke, having all the tendency to overwhelm the Dutch power. Another body of practical men inquired into the advantage of some special trades, among which the French and east India trade was found ruinous, as absorbing money and bullion, and giving in its stead but wines or spices. To these at a later date acceded the fear of Irish competition in the matter of wool. This pessimistic series of writers begins with S. Fortrey's *England's Interest and Improvement*, 1663; the author of *Britannia Languens*, 1680, and J. Pollexfen, *England and East India Inconsistent in Their Manufactures*, 1697, were its foremost champions. The commercial treaty with France in 1713 was a new matter of complaint. In the *British Merchant*, all the arguments against the underbalance are restated by Sir Theodore Janssen in his *General Maxims in Trade*, 1713, and by Joshua Gee, who afterwards put forward his views in *The Trade and Navigation of Great Britain consider'd*, 1729. 'His writings', says Hume, 'struck the nation with an universal panic, when they saw it plainly demonstrated that the balance was against them for so considerable a sum as must leave them without a single shilling in five or six years.' Nevertheless, the creed of the balance of trade was shared not only by Cantillon and Sir J. Steuart (book ii, ch. xv), but even by freetraders like Thomas Gordon, *The Nature and Weight of the Taxes of the Nation*, 1722, Vanderlint, *Money answers All Things*, 1734, and the author of *An Essay on the Causes of the Decline of Foreign Trade*, 1743.

For some time, however, the belief in the doctrine had been shaken, partly by traders whose interest it was to refute its postulates, partly by the impossibility of giving the exact statistical statement of the balance, partly by the doubts raised by superior thinkers. One of the first, it seems, was the author of *Free Ports, the Nature and Necessitie of them Stated*, B. W., 1652:

> All consultations whatsoever about trade, if free ports bee not opened and this wholesale or general trade bee not incouraged, do still but terminate in som advice or other about regulating our consumption; and have no other good at farthest, but preventional, that our Ballance of Import exceed not our Export: which to confine ourselves to alone, is, on the other side a cours to short, as it will neither serv to rais the Strength of this Nation in shipping, or to Govern the Exchange abroad (p. 8).

But the first thorough refutation was given by Nicholas Barbon in 1690 and 1696, and his influence is to be traced in the writings of Sir Dudley North (*Discourses of Trade*, 1691 edn), who calls, evidently in reference to it, the balance of trade one of the current 'politick conceits in trade; most of which Time and better Judgment hath disbanded'. The increase of manufactures had in opposition to the former opinion that 'trade was the source of national riches' made way to the doctrine that the employment of population and labour was the primitive enriching power. 'Land and labour', says therefore John Bellers, 'are the foundations of riches, and the fewer Idle Hands we have the faster we increase in value; and spending less than we raise is a much greater certainty of growing Rich than any computations that can be made from our Exportation and Importation' (*Essays about the Poor*, 1699, p. 12). These views, though far more mingled with mercantilist beliefs, were upheld by the author of *The Advantages of the East India Trade to England*

consider'd (1701 and 1720), who pointed out, that the only rule of foreign trade should be 'to get a greater for a less value', and by Defoe, who while refuting the authors of the *British Merchant*, declared himself to be 'a profess'd opposer of all fortuitous calculations, making estimates by guess work of the Quantities and Value of any Trade or Exportation' (*A Plan of the English commerce*, 1728; 2nd edn, 1737, p. 232). This confession and the doubts raised by Bishop Berkeley in his *Querist* (1735, Queries 555, 556), whether the rule of the balance of trade held always true, and whether it admitted not of exceptions, were indeed nothing new. For even Davenant, originally much devoted to these estimates (*Of the use of Political Arithmetic*, 1698, Works, vol. i, pp. 146–8), declared himself afterwards convinced that they were inaccurate for many important trades (*A Report to the Commissioners*, 1712, Works, vol. v, p. 382). Sir Josiah Child also stated, as Berkeley did, that by means of smuggling, and furthermore in the case of countries whose income was consumed by absentees, like Ireland, exports could exceed imports without enriching the people (*A new Discourse of Trade*, 1690, ch. ix). The doubts which all these expressions of opinion fostered, paved the way for the overthrow of the system. This was accelerated by the flourishing state of English trade, which continued to prosper through the 18th century notwithstanding all the predictions of evil expressed by the balance-of-trade theorists.

The successful onslaught on the system made by Hume in his Essays (1752) is now a matter of history. In these he restated Barbon's assertion that an equivalent must be paid in an export for every import received. Hume's refutation of the balance-of-trade theory had a considerable influence on the free trade doctrines of the physiocrats and also upon Adam Smith. The latter, like Barbon, controverted the theory on this subject which was laid down by Mun and by Locke. Adam Smith also, in the preference he gave to the home trade, and in his opposition to the mercantilist views, shows an inclination to incredulity in relation to the theory of foreign trade. The manner in which Adam Smith thus placed himself in opposition to the commonly-accepted opinions of his time explains the fact that his criticism of the theory of foreign trade obtained, when it first appeared, comparatively few adherents. Even Pitt, while proving the success of his policy by the growth of exports, said, when the authority of Adam Smith was quoted against him, that he considered 'that great author, though always ingenious, sometimes injudicious' (*Parliamentary History*, xxxiii, 562–3). The questioning, however, as to the complete applicability of the theory gradually extended as the 18th century waned. After the successful peace of Paris in 1763 the fear of a drain of specie began to spread in consequence of the growth of indebtedness to foreigners; and though the balance of trade seemed favourable, new doubts were expressed whether the values stated of the goods exported were accurate (*The Present State of the Nation*, 1769; by W. Knox, secretary to George Grenville, pp. 65–7). The observations of Burke on this occasion, though professedly designed to prove the balance to be favourable, are very acute. Though not allowing the statement as to the certificated goods for re-exportation to admit of error, he concedes the possibility for free goods, exported without drawback and bounty; he remembers that the costs of freight and the profits of the merchant are not taken into account, that in the balance of the Irish and West India trades import and export both refer to one nation, and he ridicules those who held that the foreign imports were a loss without even considering that part of it which enters into production (see Edmund Burke, *Observations on a late State of the Nation*, 1769, pp. 34–8. Also his *Letters on a Regicide Peace*, 1796, Works, vol. iv. p. 554). The refutation of the original theory of the balance of trade is justly ascribed to Adam Smith, and his predecessors in England, of whose principal works some notice has been given here. The work of Adam Smith was completed by Ricardo in his theory of international trade which has hitherto been the special domain of English economics.

Balanced Budget Multiplier

M. H. Peston

The balanced budget multiplier theorem is concerned with changes in aggregate demand consequent on simultaneous and equal changes in government expenditure and taxation. The essence of the theorem is that the expansionary effect of the former exceeds the contractionary effects of the latter. Thus the net effect is positive rather than zero which the commonsense of pre-Keynesian economics suggested. In other words, a tax-financed increase in public expenditure would be expansionary rather than neutral.

The theorem in its original form had a further remarkable characteristic. It was proved that the value of the balanced budget multiplier was not merely positive, but was precisely equal to unity. This appeared to be a rare example within economics of something rather less rare in natural science, namely the possibility of deriving a precise empirical magnitude from theoretical reasoning. Merely postulating, within the closed economy, a marginal propensity to consume and a marginal tax rate, both between zero and unity, led to a balanced budget multiplier of one.

The point may be seen most clearly by examining the multiplier in its form as an infinite series. The effect on aggregate demand of a unit increase in government expenditure on domestically produced goods and services is given by the series:

$$1 + c + c^2 + \cdots$$

The effect of a unit increase in income tax revenue is given by the series:

$$c + c^2 + \cdots$$

(where c in both cases is the marginal propensity to consume).

If the latter is deducted from the former, a value of unity follows for the net effect on aggregate demand.

It can be shown that if the increase in tax occurs because of an increase in the marginal tax rate, the same result follows. In addition, if imports are a function of consumption so that c is interpreted as the marginal propensity to consume domestically produced goods and services, once again the balanced budget multiplier is unity.

This, of course, immediately allows the examination of cases in which the balanced budget multiplier is positive but different from one, and even cases in which it is negative. Consider, for example, a unit increase in government expenditure, only a fraction b of which is spent on domestically produced goods and services, the remainder going on imports. The initial sequence above will then be multiplied by a number b lying between zero and one. In that sequence c will be the propensity to spend on domestic output (i.e., it is net of imports). Consider also an increase in taxation levied initially on households who spend on domestically produced goods and services a different fraction of the change in their income from the community at large. The second sequence above can then be rewritten as follows:

$$c' + c'c + c'c^2$$

where c' differs from c and is the initial decline in consumption.

The net effect on aggregate income will then be $(b - c')\ (1 + c + c^2 + \ldots)$. This equals $(b - c')/(1 - c)$. It is obvious that if c' is greater than b, the balanced budget multiplier is negative. If c' is sufficiently smaller than c, the balanced budget multiplier will be greater than unity.

Also, of course, the multiplier will change in value if the spending propensities relevant to the later stages of the government expenditure sequence ares not the same as those in the tax sequence.

The matter may be complicated further by considering various forms of taxation. The impact effect of a unit increase in income taxation is assumed to be c. The impact effect of a unit increase in indirect taxation is assumed to be unity – that is, a switch to indirect taxation leaving total tax revenue constant is contractionary. The impact effect of a unit increase in corporate taxation will, presumably, be on investment rather

than consumption spending, unless it is passed on to households in higher prices.

All of this may be summarized by noting (i) that government expenditure may be on transfer payments or on goods and services; (ii) that the part devoted to goods and services may be further subdivided into that obtained from domestic production and that from abroad; (iii) that the tax effects depend on the nature of the taxes being levied; (iv) that the initial impact on aggregate demand of a tax increase depends on who it is levied on, relevant distinctions being between firms and households, and different types of households; and (v) that different categories of tax payers have different propensities to import. It follows that a tax-financed increase in government expenditure cannot be predicted without detailed consideration of the nature of the expenditure and the taxation. What is important, however, is that there is no presumption that a balanced budget is neutral with respect to aggregate demand. This itself is another way of putting the fundamental theorem of fiscal policy: fiscal stance is not measured correctly by the difference between public expenditure and tax revenue.

The history of the balanced budget multiplier is of some interest. The theorem was originally attributed to Haavelmo (1945). It is apparent that a prior claim to publication must go to Gelting (1941). An important early contribution was that of Wallich (1944). (Others have claimed to have known of the theorem and even to have written it down without publishing it, but that is not at all the same thing.) Important generalizations are attributable to Turvey (1953) and Peston and Baumol (1955).

See Also

► Multiplier Analysis

Bibliography

Gelting, J. 1941. Nogle Bemaerkninger om Finansieringen af offentlig Virksomhed. *Nationalökonomisk Tidsskrift* 79(5): 293–299.

Haavelmo, T. 1945. Multiplier effects of a balanced budget. *Econometrica* 13(October): 311–318.

Peston, M.H., and W.J. Baumol. 1955. More on the multiplier effects of a balanced budget. *American Economic Review* 45(March): 140–148.

Turvey, R. 1953. Some notes on multiplier theory. *American Economic Review* 43: 275–295.

Wallich, H.C. 1944. Income-generating effects of a balanced budget. *Quarterly Journal of Economics* 59(November): 78–91.

Balanced Growth

Jonathan Temple

Abstract

'Balanced growth' has at least two different meanings in economics. In macroeconomics, balanced growth occurs when output and the capital stock grow at the same rate. This growth path can rationalize the long-run stability of real interest rates, but its existence requires strong assumptions. In development economics, balanced growth refers to the simultaneous, coordinated expansion of several sectors. The usual arguments for this development strategy rely on scale economies, so that the productivity and profitability of individual firms may depend on market size. The article reviews the balanced growth debate and the extent to which it has influenced development policies.

Keywords

Balanced growth; Big Push; Coordination failures; Development strategies; Economic geography; Economies of scale; Hirschman, A.; Increasing returns; Industrialization; Krugman, P.; Multiple equilibria; Nurkse, R.; Pecuniary external economies; Rosenstein-Rodan, P.; Scitovsky, T.; Solow, R.; Swan, T.; Unbalanced growth

JEL Classifications

O4

In macroeconomics, 'balanced growth' refers to classes of equilibrium growth paths, while in development economics the term refers to a particular development strategy.

These two uses of the term are clearly distinct, and each is discussed in turn.

The concept of a balanced growth path is a central element of macroeconomics.

It refers to an equilibrium in which major aggregates, usually but not exclusively output and the capital stock, grow at the same rate over time, and the real interest rate is constant. Most textbook growth models are constructed in a way that delivers this outcome. This is motivated partly by theoretical convenience but also by historical observation. The conventional wisdom is that real interest rates and the capital-output ratio are surprisingly stable over long spans of time, at least in developed countries.

Balanced growth is not an inevitable property of growth models. It was not until the publication of classic papers by Solow (1956) and Swan (1956) that economists saw how a balanced growth path might arise from relatively appealing assumptions. The key insight is that a stable equilibrium path requires the possibility of substitution between capital and labour. The Solow–Swan model has subsequently underpinned much empirical work on economic growth, and has also influenced short-run macroeconomics.

The existence of a balanced growth path requires strong assumptions. The usual derivation assumes that aggregate output can be written as a function of the total inputs of capital and labour, with diminishing returns to each input and constant returns to scale overall. In addition to the conditions needed for aggregation, either the production function should be Cobb–Douglas, or technical progress should be restricted to the labour-augmenting type. In other words, when technology advances, it should be 'as if' the economy had more labour than before, and not 'as if' it had more capital.

Because these assumptions are strong, any use of balanced growth to rationalize the data tends to create new puzzles. For example, why should technical progress be exclusively labour-augmenting, as stability of real interest rates would require? Acemoglu (2003) has examined this question using an incentives-based model of technical change, but in general balanced growth seems a less than inevitable outcome of a real-world growth process. The picture is even more complicated when there are multiple sectors, whether differentiated as capital and consumer goods, or as different types of final goods. As might be expected, where multiple sectors are present, the conditions needed for balanced growth become even stricter. Greenwood et al. (1997) and Kongsamut et al. (2001) are two useful references on multi-sector growth models.

None of this is to deny that balanced growth is a useful concept. The idea plays an important role in teaching and research in macroeconomics because of its simplicity and explanatory power. As with all organizing frameworks, however, it is sensible to be aware of its limitations and the possibilities that lie outside it.

In macroeconomics, balanced growth is usually associated with constant returns to scale. For most development economists, the term is more strongly associated with increasing returns and a debate that began with Rosenstein-Rodan (1943). He argued that the post-war industrialization of eastern and south-eastern Europe would require coordinated investments across several industries. The idea is that expansion of different sectors is complementary, because an increase in the output of one sector increases the size of the market for others. A sector that expands on its own may make a loss but, if many sectors expand at once, they can each make a profit. This tends to imply the need for coordinated expansion, or a 'Big Push', and potentially justifies a role for state intervention or development planning. Another influential contribution by Nurkse (1953) made similar points, giving more emphasis to the links between market size and the incentives to accumulate capital.

In Rosenstein-Rodan's paper the argument is set out informally, and with many digressions. But the central point will have a familiar ring to students of modern game theory and the literature on coordination failures. Essentially, Rosenstein-Rodan was setting out assumptions that might give rise to multiple equilibria in levels of development. Papers by Fleming (1955) and Scitovsky

(1954) further clarified some of the necessary assumptions. Fleming emphasized the importance of Rosenstein-Rodan's assumption that the industrializing sectors can draw on labour from other sectors without forcing up wages. Scitovsky noted that the proponents of balanced growth appeared to see externalities everywhere, but under perfect competition, external effects that are mediated through markets ('pecuniary external economies') do not preclude Pareto efficiency. This result hints at the importance of scale economies to the balanced growth hypothesis, since then market size can influence unit costs, and Scitovsky's logic no longer applies.

The key ideas of the balanced growth hypothesis were formalized in a much-admired paper by Murphy et al. (1989). In their multi-sector model, firms in each sector use constant returns-to-scale technologies, but one firm in each sector also has access to an increasing returns-to-scale technology. This technology will be profitable to operate only given a sufficiently large market. The structure of the model, with a competitive fringe of small-scale producers, ensures that wages are independent of labour demand in the industrializing sectors. The model yields multiple equilibria that can be Pareto-ranked.

The assumptions needed for multiplicity are more complicated than earlier authors believed, however. For example, increasing returns and an elastic supply of labour are not sufficient in themselves to generate multiple equilibria. Consider an equilibrium in which no sectors have industrialized (meaning that none is using the increasing returns-to-scale technique). If a single firm then adopts the modern technique and makes a loss, this will reduce rather than increase the size of the market for other sectors, so the necessary complementarity is absent. For multiple equilibria to arise, the industrializing firm must somehow raise the size of the market for other sectors, even though it makes a loss when acting alone. In one of the models considered by Murphy et al. (1989), this is achieved by an extra assumption, namely, that industrializing firms must pay higher wages than other firms.

Although the balanced growth hypothesis has been widely discussed, it has a number of limitations. The ideas are difficult to test empirically. From a purely theoretical point of view, the argument does not generalize straightforwardly to open economies. If firms can sell their output abroad, the role of domestic market size appears much less important. The balanced growth hypothesis then requires a more complex story, perhaps one in which firms are especially reliant on domestic markets in the early stages of their development.

The ideas have also been criticized on other grounds. The most prominent sceptic was Hirschman (1958), who argued that simultaneous, coordinated investment asked too much of developing countries. He regarded growth as a necessarily unbalanced dynamic process, in which successive disequilibria create the conditions for development in other sectors. Unbalanced growth could occur either through forward and backward linkages to downstream and upstream industries or by drawing out latent capacities needed for growth, such as the application of entrepreneurial skills.

Importantly, this process is seen as too complex and unpredictable to lend itself readily to a government-inspired 'Big Push', partly because governments may lack the relevant information, and partly because simultaneous investment would place too many demands on limited organizational resources. Hirschman (1958, pp. 53–4) summarized his objections by saying: 'if a country were ready to apply the doctrine of balanced growth, then it would not be underdeveloped in the first place'.

But his preferred vision has echoes of the balanced growth doctrine in its appeal to complementarities and increasing returns; Krugman (1995) discusses this point in more detail. Arguably it is not so much the assumptions that differ, but the view of equilibrium selection. One interpretation of Hirschman's critique is that the multiplicity of equilibria is illusory, because the earlier authors had missed out relevant state variables.

In practice, balanced growth ideas have had less influence on development strategies than a more general commitment to state-led industrialization and import substitution. A perceived need for balanced growth may have motivated some attempts at indicative planning, but state interventions have usually tried to focus on particular sectors rather than attempting the more ambitious

task of simultaneous expansion across many industries. The reasons for this are likely to be complex, including uncertainty over which sectors should be encouraged to expand, and the lack of obvious ways to coordinate this without direct state control. In the academic literature, the difficulty of testing the main ideas has been another factor limiting their influence.

For reasons like these, the balanced growth hypothesis is currently at the margins of development thinking and policy advice. The ideas are still interesting, however, and their neglect is partly due to the accidents of intellectual history. Formalizing Rosenstein-Rodan's original insights proved a difficult task. The reasons for this are discussed in Krugman (1995) as part of an illuminating account of the balanced growth debate and the role of formal models. He shows the continuing relevance of the main ideas to economic geography and regional science, and his book can be highly recommended to anyone interested in balanced growth, or the methods of modern economics more generally. Another useful reference is the special issue of the *Journal of Development Economics* on increasing returns and economic development (April 1996).

See Also

- ▶ Development Economics
- ▶ Growth Models, Multisector
- ▶ Linkages
- ▶ New Economic Geography
- ▶ Poverty Traps
- ▶ Regional Development, Geography of
- ▶ Structural Change

Bibliography

Acemoglu, D. 2003. Labor- and capital-augmenting technical change. *Journal of the European Economic Association* 11: 1–37.

Fleming, M. 1955. External economies and the doctrine of balanced growth. *Economic Journal* 65: 241–256.

Greenwood, J., Z. Hercowitz, and P. Krusell. 1997. Long-run implications of investment-specific technological change. *American Economic Review* 87: 342–362.

Hirschman, A. 1958. *The strategy of economic development*. New Haven: Yale University Press.

Kongsamut, P., S. Rebelo, and D. Xie. 2001. Beyond balanced growth. *Review of Economic Studies* 68: 869–882.

Krugman, P. 1995. *Development, geography, and economic theory*. Cambridge, MA/London: MIT Press.

Murphy, K., A. Shleifer, and R. Vishny. 1989. Industrialization and the big push. *Journal of Political Economy* 97: 1003–1026.

Nurkse, R. 1953. *Problems of capital gormation in underdeveloped countries*. Oxford: Basil Blackwell.

Rosenstein-Rodan, P. 1943. Problems of industrialisation of eastern and south-eastern Europe. *Economic Journal* 53: 202–211.

Scitovsky, T. 1954. Two concepts of external economies. *Journal of Political Economy* 62: 143–151.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.

Swan, T. 1956. Economic growth and capital accumulation. *The Economic Record* 32: 334–361.

Balassa, Béla (1928–1991)

Jaime de Melo and Carl F. Christ

Keywords

Balassa, B.; Bhagwati, J.; Chenery, H.; Christ, C.; Effective rate of protection; Exchange rate regimes; Import substitution; INSTITUTE for International Economics; International trade; Fischer, S.; Johns Hopkins University; Krueger, A.; Melo, J. de; Noland, M.; Protection; Purchasing power parity; Revealed comparative advantage; Summers, L.; Structural adjustment lending; Tariffs; World Bank

JEL Classifications

B31

Béla Balassa, holding degrees in law and economics, left his native Hungary when the Soviet tanks put down the 1956 revolution. In 1959 he received his Ph.D. in economics from Yale. From 1966 until his death in 1991 he was professor of economics at Johns Hopkins and a consultant to the World Bank. Influenced by events in his youth,

Béla held a deep lifelong belief in political and economic freedom.

At the World Bank, Béla was very active as Research Advisor to the Vice- President for Research, first to Hollis Chenery, then to his successors, Anne Krueger, Stanley Fischer and Larry Summers. He held this position until his death, and those of us who were then at the Bank will remember him as the Bank's most influential economic advisor during his 25 years involvement at the institution. His commitment to economic policy was extended by his involvement in his later years at the Institute for International Economics, where he wrote on trade policy issues of developed countries, notably on Japan (Balassa and Noland 1988).

Béla was among the most prolific international trade economists of his generation, contributing several books that are still widely cited. Early in his career he made several lasting contributions, among which was his famous paper on purchasing power parity (1964) in which he used a Ricardian model to show that a country's real exchange rate would appreciate as its productivity gap narrowed. Béla also made lasting contributions to the theory of economic integration (1962) and to empirical methods, proposing a measure of 'revealed comparative advantage' and ways to measure rates of effective protection (1965).

As research advisor of the Development Research Center at the World Bank, Béla fulfilled many roles during the three days a week he spent there. Whereas most members at the centre would devote most of their time to research, in addition to his highly productive research activities Béla participated very actively in the Bank's policy dialogue, commenting on the vast majority of country reports, and invariably on all those that contained advice on trade policies. In those days trade policy was a major issue in virtually all countries. Then, import-substitution policies supported by highly restrictive trade regimes were the rule. With a handful of trade economists, including Jagdish Bhagwati and Anne Krueger, Béla would tirelessly recommend a simplification of the trade regime, moderate protection of industrial activities supported by uniform tariffs, a removal of quantitative restrictions, and a unification of the then prevailing multiple exchange rate regimes.

Béla's advice on trade policy was supported by his research carried out under the Bank's auspices. He directed and edited an influential book that examined the trade regimes of several countries in Latin America and East Asia, documenting systematically the patterns of effective rates of protection in these countries (Balassa and Associates 1971).

Béla's research output was not only prolific but also timely. His ability to be the first to deliver relevant research on the policy issue of the day was uncanny. In the late 1970s, when developing countries were hit by oil, commodity and interest rate shocks, Béla was the first to implement a useful decomposition formula to assess the extent of purchasing power loss. Later, when the Bank launched structural adjustment lending activities and wanted to assess performance of countries having received adjustment loans, Béla again delivered the first assessment of adjustment lending.

Béla's work capacity was legendary. Despite his influential research and his sage and realistic policy advising at the World Bank, which left him only two days a week for Johns Hopkins, his contribution to teaching, thesis supervision and academic governance at Hopkins was enormous. He taught most of the courses in international and development economics. He supervised more students than almost anyone else, and he responded to their papers and thesis drafts almost instantly with demanding but constructive comments. For ten years he was an elected and reelected member of the faculty governing council. As chair of the faculty budget committee, he persuaded the university to reverse the decline that had been permitted to occur in the real value of its tuition charge, its faculty compensation levels and its academic expenditures.

Besides all this, Béla was an informed lover of art, opera, French literature and food (his guide to Paris restaurants was prized), and he always made time for his friends and for his family.

See Also

► Purchasing Power Parity

Selected Works

1962. *The theory of economic integration*. Homewood: Richard D. Irwin Inc.
1964. The purchasing power parity doctrine: A reappraisal. *Journal of Political Economy* 72:584–596.
1965. Tariff protection in industrial countries: An evaluation. *Journal of Political Economy* 73:573–594.
1971. (With Associates.) *The structure of protection in developing countries*. Baltimore: Johns Hopkins University Press.
1988. (With M. Noland.) *Japan and the world economy*. Washington, DC: Institute for International Economics.

Balch, Emily Greene (1867–1961)

Robert W. Dimand

Abstract

The American economist, sociologist and pacifist Emily Greene Balch shared the Nobel Peace Prize in 1946 for the same anti-war activism for which she was not reappointed as a full professor of economics and sociology at Wellesley College in 1918. She was also notable as a defender of the economic, social and cultural benefits of ethnically diverse immigration, at a time when many economists wished to restrict immigration.

Keywords

Academic freedom; Economics in USA; Immigration; Pacifism; Women in economics

JEL Classifications

B31

The American economist and sociologist Emily Greene Balch shared the Nobel Peace Prize in 1946 for the same anti-war activism for which she lost her professorship at Wellesley College in 1918.

Emily Greene Balch was born in Jamaica, Plain, Massachusetts, in 1867, the second of six surviving children of a former school-teacher and of a lawyer who, after graduating from Harvard in 1859, became secretary to the abolitionist Senator Charles Sumner. She was part of Bryn Mawr's first graduating class in 1889, having studied with the sociologist Franklin Giddings and with Woodrow Wilson, then assistant professor of history and government at Bryn Mawr and a founding member of the American Economic Association's council. Balch spent 1890–91 at the Sorbonne as the first winner of the Bryn Mawr Fellowship for European Study, researching and writing her monograph on *Public Assistance of the Poor in France*, which was published by the American Economic Association in 1893. Balch studied economics under economic historian William Ashley for a semester at the Harvard Annex (later Radcliffe) in 1893, while engaging in social work as head of Denison House, a Boston settlement house (a place of refuge and support for the poor) patterned on Jane Addams's Hull House in Chicago. After attending the 1894 national convention of the American Federation of Labor as a delegate from Boston's Central Labor Union, she took courses in economic theory and sociology at the University of Chicago for a quarter in 1895, and spent 1895–96 at the University of Berlin, attending the seminars of public finance specialist Adolf Wagner and historical economist Gustav Schmoller – and also the International Socialist Workers and Trade Union Congress in London in 1896.

Balch returned from Germany on the same ship as Katharine Coman (later author of the lead article in the inaugural issue of the *American Economic Review*), who was then the only economics teacher at Wellesley College. Coman invited Balch to join Wellesley, at first in a half-time job grading papers but teaching economics courses from the second semester (and from 1900 also

Wellesley's first courses in sociology). At Wellesley, Balch taught courses on immigration, labour problems, the history of socialism, social pathology, consumption, the economic role of women, introductory economics, sociology, statistics and economic history, while also serving on the Massachusetts Factory Inspection Commission and Boston's City Planning Board and chairing the Massachusetts Minimum Wage Commission, which drafted the country's first minimum wage law. She was promoted to associate professor in 1903, and in 1913 received a five-year contract as full professor and head of the Department of Economics and Sociology, in succession to the ailing Coman. While many American economists of the day were active in the Immigration Restriction League, Emily Balch upheld the social, cultural and economic benefits of free immigration from diverse sources. Her major work, *Our Slavic Fellow Citizens* (1910), resulted from a sabbatical in Austria-Hungary in 1904–1905 visiting sources of Slavic immigration to the USA and unpaid leave in 1905–1906 visiting centers of Slavic immigration in the USA.

A pacifist since the Spanish-American War (and in 1921 a convert from Unitarianism to Quakerism), Balch was active in the International Congress of Women at The Hague in 1915, urging a conference of neutral nations to offer mediation to end the First World War, and spent several months with the International Committee on Mediation in Stockholm in 1916, as well as lobbying her former teacher, President Wilson. Balch wrote for *The Nation* during a sabbatical (1916–17) and unpaid leave (1917–18), opposing conscription and defending civil liberties, including those of conscientious objectors and the foreign born, and published *Approaches to the Great Settlement* (1918) on how to end the war. Balch's contract expired in 1918, at a time when many American universities and colleges from Columbia to Nebraska were dismissing anti-war (or insufficiently pro-war) faculty. Ostensibly because of the length of her leave of absence, Wellesley's trustees narrowly voted the next year against reappointment, despite the protests of Wellesley's president Ellen Pendleton, the alumnae trustees, and Balch's department. Balch regretted having 'overstrained the habitual liberality' of Wellesley and never uttered recrimination, but, although Wellesley College later made amends (inviting Balch to give the Armistice Day address in 1935) and although she lived in Wellesley, Massachusetts until her ninetieth year, Balch left her papers to Swarthmore College.

In 1919, Balch became the founding secretary-treasurer of the Women's International League for Peace and Freedom (WILPF), succeeding her close friend Jane Addams as president of the American section in 1931 (the year Addams won the Nobel Peace Prize) and as honorary international president in 1937. She was influential in urging the removal of US Marines from Haiti (Balch (ed.) *Occupied Haiti*, 1927). Appalled by Nazi aggression and persecution of Jews, and in view of the Pearl Harbor attack, Balch supported US entry into the Second World War as the lesser evil, but stayed in the WILPF to defend the rights of Japanese-Americans and conscientious objectors, urge the admission of refugees and oppose Allied demands for unconditional surrender. She won the Nobel Peace Prize in 1946 (jointly with John R. Mott of the Student Christian Movement), the third woman to win the Peace Prize and the first American economist to win a Nobel Prize. Balch died in January 1961. For further reading, see Balch (1972) and Randall (1964).

See Also

► Immigration and the City

Selected Works

1893. Public assistance of the poor in France. *Publications of the American Economic Association*, First Series 8(4–5): 1–180.
1910. *Our Slavic fellow citizens*. New York: N.Y. Charities Publication Committee.
1918. *Approaches to the great settlement*. New York: B.W. Heubsch.

1927. *Occupied Haiti*. New York: The Writers Publishing Company.
1972. *Beyond nationalism: The social thought of Emily Greene Balch*, ed. M.M. Randall. New York: Twayne Publishers.

Bibliography

Randall, M.M. 1964. *Improper Bostonian: Emily Greene Balch, Nobel peace laureate, 1946*. New York: Twayne Publishers.

Balogh, Thomas (1905–1985)

P. Streeten

Balogh was one of that influential group of exiled Hungarian economists, for whose ambitions and talents Hungary was too small and poor. Experience of the power politics of the 1930s, as seen from a Hungary dominated by Germany, equipped him well to understand the adjustments of post-imperial Britain to a world in which power had ebbed away from her. Under the influence of the banker O.T. Falk, also the originator of many of Keynes's ideas, Balogh was converted from an anti-inflationary creed to his fierce hostility to dear money and deflationary policies. His *Studies in Financial Organization* (1947) combines a passion for reform with skilful command of intricate detail. After the war, Balogh turned his attention to the problems of the underdeveloped countries. As adviser to the Food and Agriculture Organization of the United Nations (1957–9) he transformed an afforestation project into a series of ambitious development plans of the countries round the Mediterranean.

After Harold Wilson resigned from the Cabinet in 1951 he came into close touch with Balogh. One of Balogh's lines of argument was that a Labour government should be committed to a policy of faster growth, sustained by a strong incomes policy and supported by more state intervention in industry and foreign exchange controls. After the Labour victory of 1964 Balogh was brought into the Cabinet Office as adviser on economic affairs, with special reference to external economic policy. After three-and-a-halfyears of service in Number 10 Downing Street he was made a life peer and returned to the University of Oxford.

Although often labelled an extreme left-wing economist, he challenged many cherished socialist clichés. Having moved gradually to the left (he had been a follower of Horthy, later a liberal, and did not become a socialist until the war), he believed in linking together like-minded nations, both rich and poor, which would build up their jointly planned economies behind protective barriers, on the basis of high investment, modernization and fair shares. He favoured central planning and controls because he believed that they alone could secure an efficient and fair allocation of resources.

He identified many problems before the bulk of the profession had turned its attention to them. Among these were the scale of German rearmament in the 1930s, the need for exchange control during the war, the dollar problem after the war, the importance of an incomes policy based on a social consensus, the need for international coordination of demand management, the role of rural education and agriculture in development, the content and style of higher education in Africa, and the need for professional expertise in the Civil Service.

Superficially, his views seem full of contradictions, such as his advocacy of administrative controls while denouncing administrators. Yet there is a unity of vision behind these paradoxes, often guided more by intuition than by formal analysis.

Selected Works

1947. *Studies in financial organization*. Cambridge: Cambridge University Press.
1949. *The dollar crisis*. Oxford: Basil Blackwell.
1963. *Unequal partners*. Oxford: Basil Blackwell.
1964. *The economic impact of monetary and commercial institutions of a European origin in Africa*. Cairo: National Bank of Egypt.

1973. *Facts and fancy in international economic relations*. Oxford: Pergamon.
1983. *The irrelevance of conventional economics*. London: Weidenfeld & Nicolson.

Bandit Problems

Dirk Bergemann and Juuso Välimäki

Abstract

The multi-armed bandit problem is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. This classic problem has received much attention in economics as it concisely models the tradeoff between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff).

Keywords

Asset pricing; Bandit problems; Branching bandit problem; Continuous-Time models; Corporate finance; Descending price auction; Experimentation; Free-Rider problem; Gittins index th; Informational efficiency; Learning; Liquidity; Markov equilibria; Matching markets; Moral hazard; Noise trader; Probability distribution; Product differentiation; Regime switch

JEL Classifications

C72; C73; D43; D81; D82; D83; D92; G24; G31

The multi-armed bandit problem, originally described by Robbins (1952), is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. In the multi-arm bandit problem, the gambler has to decide which arm of K different slot machines to play in a sequence of trials so as to maximize his reward. This classical problem has received much attention because of the simple model it provides of the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff). Each choice of an arm results in an immediate random payoff, but the process determining these payoffs evolves during the play of the bandit. The distinguishing feature of bandit problems is that the distribution of returns from one arm only changes when that arm is chosen. Hence the rewards from an arm do not depend on the rewards obtained from other arms. This feature also implies that the distributions of returns do not depend explicitly on calendar time.

The bandit framework found early applications in the area of clinical trials where different treatments need to be experimented with while minimizing patient losses and in adaptive routing efforts for minimizing delays in a network. In economics, experimental consumption is a leading example of an intertemporal allocation problem where the trade-off between current payoff and value of information plays a key role.

Basic Model

It is easiest to formulate the bandit problem as an infinite horizon Markov decision problem in discrete time with time index $t = 0, 1, \ldots$ At each t, the decision maker chooses amongst K arms and we denote this choice by $a_t \in \{1, \ldots, K\}$. If $a_t = k$, a random payoff x_t^k is realized and we denote the associated random variable by X_t^k. The state variable of the Markovian decision problem is given by s_t. We can then write the distribution of x_t^k as $F^k(\cdot; s_t)$. The state transition function φ depends on the choice of the arm and the realized payoff:

$$s_{t+1} = \varphi\left(x_t^k; s_t\right)$$

Let S_t denote the set of all possible states in period t. A feasible Markov policy $a = \{a_t\}_{t=0}^{\infty}$ selects an available alternative for each conceivable state s_t, that is,

$$a_t : S_t \to \{1, \ldots, K\}$$

The following two assumptions must be met for the problem to qualify as a bandit problem.

1. Payoffs are evaluated according to the discounted expected payoff criterion where the discount factor δ satisfies $0 \leq \delta < 1$.
2. The payoff from each k depends only on outcomes of periods with $a_t = k$. In other words, we can decompose the state variable s_t into K components $\left(s_t^1, \ldots, s_t^K\right)$ such that for all k:

$$\begin{array}{lll} s_{t+1}^k = s_t^k & \text{if} & a_t \neq k, \\ s_{t+1}^k = \varphi\left(s_t^k, x_t\right) & \text{if} & a_t = k, \end{array}$$

and

$$F^k(\cdot, s_t) = F^k\left(\cdot; s_t^k\right).$$

Notice that when the second assumption holds, the alternatives must be statistically independent.

It is easy to see that many situations of economic interest are special cases of the above formulation. First, it could be that $F^k(\cdot\,;\theta^k)$ is a fixed distribution with an unknown parameter θ^k. The state variable is then the posterior probability distribution on θ^k. Alternatively, $F^k(\cdot\,;s^k)$ could denote the random yield per period from a resource k after extracting s^k units.

The value function $V(s_0)$ of the bandit problem can be written as follows. Let $X^k\left(s_t^k\right)$ denote the random variable with distribution $F^k\left(\cdot; s_t^k\right)$. Then the problem of finding an optimal allocation policy is the solution to the following intertemporal optimization problem:

$$V(s_0) = \sup_a \left\{ \mathbb{E} \sum_{t=0}^{\infty} \delta^t X^{a_t}\left(s_t^{a_t}\right) \right\}.$$

The celebrated index theorem due to Gittins and Jones (1974) transforms the problem of finding the optimal policy into a collection of k stopping problems. For each alternative k, we calculate the following index $\gamma^k\left(s_t^k\right)$, which depends only on the state variable of alternative k:

$$m^t\left(s_t^k\right) = \sup_{\tau} \left\{ \frac{\mathbb{E}\sum_{u=t}^{\tau} \delta^t X^k\left(s_u^k\right)}{\mathbb{E}\sum_{u=t}^{\tau} \delta^t} \right\}, \qquad (1)$$

where τ is a stopping time with respect to $\left\{s_t^k\right\}$. The idea is to find for each k the stopping time τ that results in the highest discounted expected return per discounted expected number of periods in operation. The Gittins index theorem then states that the optimal way of choosing arms in a bandit problem is to select in each period the arm with the highest Gittins index, $m^k\left(s_t^k\right)$, as defined by (1).

Theorem 1 Gittins-Jones (1974) *The optimal policy satisfies $a_t = k$ for some k such that*

$$m^k\left(s_t^k\right) \geq m^j\left(s_t^j\right) \text{for all } j \in \{1, \ldots, K\}.$$

To understand the economic intuition behind this theorem, consider the following variation on the original problem. This reasoning follows the lines suggested in Weber (1992). The arms are owned and operated by separate riskneutral agents. The owner can rent a single arm at a time to the operators and there is a competitive market of potential operators. As time is discounted, it is clearly optimal to obtain high rental incomes in early periods of the model. The rental market is operated as a descending price auction where the fee for operating an arbitrary arm is lowered until an operator accepts the price. At the accepted price, the operator is allowed to operate the arm as long as it is profitable. Since the market for operators is competitive, the price is such that, under an optimal stopping rule, the operator breaks even. Hence the highest acceptable price for arm k is the Gittins index $m^k\left(s_t^k\right)$, and the operator operates the arm until its Gittins index falls below the price, that is, its original Gittins index. Once an arm is abandoned, the process of lowering the price offer is restarted. Since the operators get zero surplus and they are operating under optimal rules, this method of allocating arms results in the maximal surplus to the owner and thus the largest sum of expected discounted payoffs.

The optimality of the index policy reduces the dimensionality of the optimization problem. It says that the original K-dimensional problem can

be split into K independent components, and then be knitted together after the solutions of the indices for the individual problems have been computed, as in Eq. (1). In particular, in each period of time, at most one index has to be re-evaluated; the other indices remain frozen.

The multi-armed bandit problem and many variations are presented in detail in Gittins (1989) and Berry and Fristedt (1985). An alternative proof of the main theorem, based on dynamic programming can be found in Whittle (1982). The basic idea is to find for every arm a retirement value M_t^k, and then to choose in every period the arm with the highest retirement value. Formally, for every arm k and retirement value M, we can compute the optimal retirement policy given by:

$$V^k\left(s_t^k, M\right) \underline{\Delta} \max\left\{\mathbb{E}\left[X^k\left(s_u^k\right) + \delta V^k\left(s_t^{k+1}, M\right), M\right]\right\} \tag{2}$$

The auxiliary decision problem given by (2) compares in every period the trade-off between continuation with the reward process generated by arm k or stopping with a fixed retirement value M. The index of arm k in the state s_t^k is the highest retirement value at which the decision is just indifferent between continuing with arm k or retiring with $M = M\left(s_t^k\right)$:

$$M^k\left(s_t^k\right) = V^k\left(s_t^k, M^k\left(s_t^k\right)\right).$$

The resulting index $M^k\left(s_t^k\right)$ is equal to the discounted sum of flow index $m^k\left(s_t^k\right)$, or $M^k\left(s_t^k\right) = m^k\left(s_t^k\right)/(1-\delta)$.

Extensions

Even though it is easy to write down the formula for the Gittins index and to give it an economic inpt, it is normally impossible to obtain analytical solutions for the problem. One of the few settings where such solutions are possible is the continuoustime bandit model where the drift of a Brownian motion process is initially unknown and learned through observations of the process. Karatzas (1984) provides an analysis of this case when the volatility parameter of the process is known.

From an analytical standpoint, the key property of bandit problems is that they allow for an optimal policy that is defined in terms of indices that are calculated for the individual arms. It turns out that this property does not generalize easily beyond the bandit problem setting. One instance where such a generalization is possible is the branching bandit problem where new arms are born to replace the arm that was chosen in the previous period (see Whittle 1981).

An index characterization of the optimal allocation policy can still be obtained without the Markovian structure. Varaiya et al. (1985) give a general characterization in discrete time, and Karoui and Karatzas (1997) provide a similar result in a continuous time setting. In either case, the essential idea is that the evolution of each arm depends only on the (possibly entire) history and running time of the arm under consideration, but not on the realization nor the running time of the other arms. Banks and Sundaram (1992) show that the index characterization remains valid under some weak additional condition even if the number of indices is countable, but not necessarily finite.

On the other hand, it is well known that an index characterization is not possible when the decision maker must or can select more than a single arm at each t. Banks and Sundaram (1994) also show further that an index characterization is not possible when an extra cost must be paid to switch between arms in consecutive periods. Bergemann and Välimäki (2001) consider a stationary setting in which there is an infinite supply of *ex ante* identical arms available. Within the stationary setting, they show that an optimal policy follows the index characterization even when many arms can be selected at the same time or when a switching cost has to be paid to move from one arm to another.

Market Learning

In economics, bandit problems were first used to model search processes. The first paper that used a

one-armed bandit problem in economics is Rothschild (1974), in which a single firm is facing a market with unknown demand. The true market demand is given by a specific probability distribution over consumer valuations. However, the firm initially has a prior probability over several possible market demands. The problem for the firm is to find an optimal sequence of prices to learn more about the true demand while maximizing its expected discounted profits. In particular, Rothschild shows that *ex ante* optimal pricing rules may well end up using prices that are ex *post* suboptimal (that is, suboptimal if the true distribution were to be known). If several firms were to experiment independently in the same market, they might offer different prices in the long run. Optimal experimentation may therefore lead to price dispersion in the long run as shown formally in McLennan (1984).

In an extension of Rothschild, Keller and Rady (1999) consider the problem of the monopolist facing an unknown demand that is subject to random changes over time. In a continuous time model, they identify conditions on the probability of regime switch and discount rate under which either very low or very high intensity of experimentation is optimal. With a low-intensity policy, the tracking of the actual demand is poor and the decision maker eventually becomes trapped, in contrast with a high-intensity policy demand, which is tracked almost perfectly. Rustichini and Wolinsky (1995) examine the possibility of mis-pricing in a two-armed bandit problem when the frequency of change is small. Nonetheless, they show that it is possible that learning will cease even though the state of demand continues to change.

The choice between various research projects often takes the form of a bandit problem. In Weitzman (1979) each arm represents a distinct research project with a random prize associated with it. The issue is to characterize the optimal sequencing over time in which the projects should be undertaken. It shows that as novel projects provide an option value to the research, the optimal sequence is not necessarily the sequence of decreasing expected rewards (even when there is discounting). Roberts and Weitzman (1981) consider a richer model of choice between R&D processes.

Many-Agent Experimentation

The multi-armed bandit models have recently been used as a canonical model of experimentation in teams. In Bolton and Harris (1999) and Keller et al. (2005) a set of players choose independently between the different arms. The reward distributions are fixed, but characterized by parameters that are initially unknown to the players. The model is one of common values in the sense that all players receive independent draws from the same distribution when choosing the same arm. It is assumed that outcomes in all periods are publicly observable, and as a result a free riding problem is created. Information is a public good and each individual player would prefer to choose the current payoff maximizing arm and let other players perform costly experimentation with currently inferior arms. These papers characterize equilibrium experimentation under different assumptions on the reward distributions. In Bolton and Harris (1999) the model of uncertainty is a continuous time model with unknown drift and know variance, whereas in Keller et al. (2005) the underlying uncertainty is modelled by an unknown Poisson parameter.

Experimentation and Matching

The bandit framework has been successfully applied to learning in matching markets such as labour and consumer good markets. An early example of this is given in the job-market matching model of Jovanovic (1979), who applies a bandit problem to a competitive labour market. Suppose that a worker must choose employment in one of K firms and her (random) productivity in firm k is parametrized by a real variable θ^k. The bandit problem is then a natural framework for the study of learning about the match-specific productivities. For each k, s_0^k is then simply the prior on θ^k and s_t^k is the posterior distribution given s_0^k and x_s^k for $s < t$. Over time, a worker's productivity

in a specific job becomes known more precisely. In the event of a poor match, separation occurs in equilibrium and job turnover arises as a natural byproduct of the learning process. On the other hand, over time the likelihood of separation eventually decreases as, conditional on being still on the job, the likelihood of a good match increases. The model hence generates a number of interesting empirical implications which have since been investigated extensively. Miller (1984) enriches the above setting by allowing for a priori different occupations, and hence the sequence in which a worker is matched over time to different occupations is determined as part of the equilibrium.

Experimentation and Pricing

In a related literature, bandit problems have been taken as a starting point for the analysis of division of surplus in an uncertain environment. In the context of a differentiated product market and a labour market respectively, Bergemann and Välimäki (1996) and Felli and Harris (1996) consider a model with a single operator and a separate owner for each arm. The owners compete for the operator's services by offering rental prices. These models are interested in the efficiency and the division of the surplus resulting from the equilibrium of the model. In both models, arms are operated according to the Gittins index rule, and the resulting division of surplus leaves the owners of the arms as well as the operator with positive surpluses. In Bergemann and Välimäki (1996), the model is set in discrete time and a general model of uncertainty is considered. The authors interpret the experiment as the problem of choosing between two competing experience goods, in which both seller and buyer are uncertain about the quality of the match between the product and the preferences of the buyer. In contrast, Felli and Harris (1996) consider a continuous model with uncertainty represented by a Brownian motion and interpret the model in the context of a labour market. Both models show that, even though the models allow for a genuine sharing of the surplus, allocation decisions are surplus maximizing in all Markovian equilibria, and each competing seller receives his marginal contribution to the social surplus in the unique cautious Markovian equilibrium. Bergemann and Välimäki (2006) generalize the above efficiency and equilibrium characterization from two sellers to an arbitrary finite number of sellers in a deterministic setting. Their proof uses some of the techniques first introduced in Karoui and Karatzas (1997). On the other hand, if the market consists of many buyers and each one of them is facing the same experimentation problem, then the issue of free-riding arises again. Bergemann and Välimäki (2000) analyse a continuous time model as in Bolton and Harris (1999), but with strategic sellers. Surprisingly, the inefficiency observed in the earlier paper is now reversed and the market equilibrium displays too much information. As information is a public good, the seller has to compensate an individual buyer only for the impact his purchasing decision has on his own continuation value, and not for its impact on the change in continuation value of the remaining buyers. As experimentation leads in expectation to more differentiation, and hence less price competition, the sellers prefer more differentiation, and hence more experimentation to less. As each seller has to compensate only the individual buyers, not all buyers, the social price of the experiment is above the equilibrium price, leading to excess experimentation in equilibrium.

Experimentation in Finance

Recently, the paradigm of the bandit model has also been applied in corporate finance and asset pricing. Bergemann and Hege (1998, 2005) model a new venture or innovation as a Poisson bandit model with variable learning intensity. The investor controls the flow of funding allocated to the new project and hence the rate at which information about the new project arrives. The optimal funding decision is subject to a moral hazard problem in which the entrepreneur controls the unobservable decision to allocate the funds to the project. Hong and Rady (2002) introduce experimentation in an asset pricing model with uncertain liquidity supply. In contrast to the standard noise trader model, the strategic seller can

learn about liquidity from past prices and trading volume. This learning implies that strategic trades and market statistics such as informational efficiency are path-dependent on past market outcomes.

See Also

- ▶ Competition
- ▶ Diffusion of Technology

Bibliography

Banks, J., and R. Sundaram. 1992. Denumerable-armed bandits. *Econometrica* 60: 1071–1096.

Banks, J., and R. Sundaram. 1994. Switching costs and the Gittins index. *Econometrica* 62: 687–694.

Bergemann, D., and U. Hege. 1998. Dynamic venture capital financing, learning and moral hazard. *Journal of Banking and Finance* 22: 703–735.

Bergemann, D., and U. Hege. 2005. The financing of innovation: Learning and stopping. *RAND Journal of Economics* 36: 719–752.

Bergemann, D., and J. Välimäki. 1996. Learning and strategic pricing. *Econometrica* 64: 1125–1149.

Bergemann, D., and J. Välimäki. 2000. Experimentation in markets. *Review of Economic Studies* 67: 213–234.

Bergemann, D., and J. Välimäki. 2001. Stationary multi choice bandit problems. *Journal of Economic Dynamics and Control* 25: 1585–1594.

Bergemann, D., and J. Välimäki. 2006. Dynamic price competition. *Journal of Economic Theory* 127: 232–263.

Berry, D., and B. Fristedt. 1985. *Bandit problems*. London: Chapman and Hall.

Bolton, P., and C. Harris. 1999. Strategic experimentation. *Econometrica* 67: 349–374.

Felli, L., and C. Harris. 1996. Job matching, learning and firm-specific human capital. *Journal of Political Economy* 104: 838–868.

Gittins, J. 1989. *Allocation indices for multi-armed bandits*. London: Wiley.

Gittins, J., and D. Jones. 1974. A dynamic allocation index for the sequential allocation of experiments. In *Progress in statistics*, ed. J. Gani. Amsterdam: North-Holland.

Hong, H., and S. Rady. 2002. Strategic trading and learning about liquidity. *Journal of Financial Markets* 5: 419–450.

Jovanovic, B. 1979. Job search and the theory of turnover. *Journal of Political Economy* 87: 972–990.

Karatzas, I. 1984. Gittins indices in the dynamic allocation problem for diffusion processes. *Annals of Probability* 12: 173–192.

Karoui, N., and I. Karatzas. 1997. Synchronization and optimality for multi-armed bandit problems in continuous time. *Computational and Applied Mathematics* 16: 117–152.

Keller, G., and S. Rady. 1999. Optimal experimentation in a changing environment. *Review of Economic Studies* 66: 475–507.

Keller, G., S. Rady, and M. Cripps. 2005. Strategic experimentation with exponential bandits. *Econometrica* 73: 39–68.

McLennan, A. 1984. Price dispersion and incomplete learning in the long run. *Journal of Economic Dynamics and Control* 7: 331–347.

Miller, R. 1984. Job matching and occupational choice. *Journal of Political Economy* 92: 1086–1120.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 55: 527–535.

Roberts, K., and M. Weitzman. 1981. Funding criteria for research, development and exploration of projects. *Econometrica* 49: 1261–1288.

Rothschild, M. 1974. A two-armed bandit theory of market pricing. *Journal of Economic Theory* 9: 185–202.

Rustichini, A., and A. Wolinsky. 1995. Learning about variable demand in the long run. *Journal of Economic Dynamics and Control* 19: 1283–1292.

Varaiya, P., J. Walrand, and C. Buyukkoc. 1985. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Transactions on Automatic Control* AC-30: 426–439.

Weber, R. 1992. On the Gittins index for multi-armed bandits. *Annals of Applied Probability* 2: 1024–1033.

Weitzman, M. 1979. Optimal search for the best alternative. *Econometrica* 47: 641–654.

Whittle, P. 1981. Arm-acquiring bandits. *Annals of Probability* 9: 284–292.

Whittle, P. 1982. *Optimization over time*. Vol. 1. Chichester: Wiley.

Banfield, Thomas Charles (1800–?1882)

R. D. Collison Black

Banfield resided for some years in Germany and was tutor to the sons of King Ludwig I of Bavaria. After his return to England he lectured on political economy at Cambridge from 1844 until 1855, but in 1846, through the patronage of Sir Robert Peel, he became Secretary to the Privy Council.

His residence in Germany enabled Banfield to act as an interpreter both of its economy and its economists to English audiences. His 1845 Cambridge lectures were expressly intended to direct attention to 'principles that foreign authors have laid down'; Banfield referred mainly to the works of Hermann, Storch and Rossi, and seems to have been the first English writer to mention von Thünen. His concept of organization of industry was based on a theory of consumption starting from the proposition: 'that the satisfaction of every lower want in the scale creates a desire of a higher character. If the higher desire existed previous to the satisfaction of the primary want it becomes more intense when the latter is removed' (Banfield 1845, p. 11). Jevons quoted this approvingly in his *Theory of Political Economy*, but pointed out that satisfaction of lower wants does not create higher wants: 'it merely permits the higher want to manifest itself.'

The graduated scale of wants outlined by Banfield would then result in a corresponding graduated scale of industries. The organization of industry he thus related to the utility of the goods produced and pointed out the linkage between the demand for goods and the payments to factors of production. Banfield's theory of consumption led him to criticize the Ricardian theory of rent with its implications of a rising cost of satisfying primary wants, and to support free trade. His books on the *Industry of the Rhine* were purely factual, but remain useful as sources of historical information.

Selected Works

1843. *Six letters to Sir Robert Peel on the dangerous tendency of the theory of rent advocated by Ricardo*, by a Political Economist. London.

1845. *Four lectures on the organization of industry, being part of a course delivered in the University of Cambridge in Easter Term, 1844*. London: Richard and John E. Taylor.

1846. *Industry of the Rhine, series I, agriculture*. London: Charles Knight & Co.

1848. *Industry of the Rhine, series II, manufactures*. London: C. Cox.

1852. *Free production having freed trade! The pressure of taxation exposed in a lecture delivered in the University of Cambridge*. London: W. Ridgway.

1855. *A letter to William Brown Esq., MP, on the advantages of his proposed system of Decimal Coinage*. London: Robert Hardwicke.

Bank of England

Charles A. E. Goodhart

Abstract

The Bank of England, founded in 1694 to finance war against France, soon became Britain's largest bank. It became responsible for maintaining the gold standard and acting as lender of last resort. To do so, it had to withdraw from commercial banking. After failing to stay on gold (1931) the Bank became subservient to the Chancellor in macro-monetary policy and was nationalized in 1946. Operational independence to set interest rates in pursuit of an inflation target was restored in the 1990s, while its previous functions, notably bank supervision, debt management, and foreign exchange intervention, fell away.

Keywords

Bagehot, W.; Bank for International Settlements; Bank of England; Bank of Scotland; Bank rate; Banking crises; Banking supervision; Basel Committee on Banking Regulation and Supervisory Practices; Big Bang; Bretton Woods system; Bullion; Central bank independence; Central banking; Cost-push theory of inflation; Dutch East India Company; European Monetary System; Exchange controls; Exchange rate mechanism; Exchange rate targets; Financial intermediaries; Financial liberalization; Financial repression; Financial Services Authority; Friedman, M.; Gold

standard; Incomes policies; Inflation; Inflation targeting; London Clearing House; Medium Term Financial Strategy (UK); Monetarism; Monetary policy; Monetary targets; Natural rate of unemployment; Phillips curve; Radcliffe Report; Royal Bank of Scotland; South Sea Company; Stagflation; Thornton, H.; Trade unions; Treasury bills; Velocity of circulation

JEL Classifications
E5

The primary motivation for the establishment of the Bank of England was the need to raise funds to help the government finance the then current war against France, although the view had also developed that a bank could help to 'stabilize' financial activity in London given periodic fluctuations in the availability of currency and credit. An original proposal by William Paterson in 1693 for a government 'fund of perpetual interest' was turned down in favour of another proposal by Paterson in 1694 to establish a company known as the Governor and Company of the Bank of England, whose capital, once raised, would be lent in its entirety to the government.

An ordinary finance act, now known as the Bank of England Act (1694), stipulated that the Bank was to be established via stock subscriptions which were to be lent to the government. A governor, deputy governor and 24 directors were to be elected by stockholders (holding d500 or more of stock).

The Evolution of the Bank's Objectives and Functions, 1694–1914

Under its original charter the Bank was allowed to issue bank notes, redeemable in silver coin, as well as to trade in bills and bullion. The notes of the Bank competed with other paper media of exchange, which comprised notes issued by the Exchequer and by private financial companies. In addition, customers could maintain deposit accounts with the Bank, which were transferable to other parties via notes drawn against deposit receipts (known as accomptable notes), thus providing an early form of cheque.

An early customer of the Bank was the Royal Bank of Scotland, which made arrangements to keep cash at the Bank from its outset in 1727. Loans were extended, predominantly in the form of discounting of bills, to individuals and companies, and the Bank undertook a large amount of lending (often via overdrafts) to the Dutch East India Company and, from 1711, to the South Sea Company. The Bank also acted as a mortgage lender, although this business never took off, and ceased some years later. Finally, an important function of the Bank was the remittance of cash to Flanders and elsewhere for the wars against Louis XIV, which was facilitated through correspondent arrangements with banks in Holland.

In 1697 the renewal of the Bank's charter for another ten years involved the passage of a second Bank Act, which increased the capital of the Bank and prohibited any other banks from being chartered in England and Wales. This monopoly was strengthened at the next renewal of the Bank's charter in 1708, when any association of six or more persons was forbidden to engage in banking activity, thereby precluding the establishment of any other joint stock banks. The Bank's position as banker to the government was consolidated in 1715 when it was decided that subscriptions for government debt issues would be paid to the Bank, and further that the Bank was to manage the government debt (the Ways and Means Act). The Bank then acted as manager of the government's debts from that date until 1997.

The Bank also encouraged the use of its own notes in preference to other media of exchange by persuading the Treasury to increase the denomination of Exchequer bills. By 1725 the Bank's notes had become sufficiently widely used as to be pre-printed for the first time. Although a number of private banks had developed by 1750, both within and outside London, none competed seriously with the Bank in the issue of notes. By 1770 most London bankers had ceased to issue notes, using Bank of England notes (and cheques) to settle balances among themselves in what had become a well-developed clearing system. Furthermore, in 1775 Parliament raised the minimum

denomination for any non-Bank of England notes to one pound and, two years later, to five pounds, effectively guaranteeing the use of Bank of England notes as the dominant form of currency. Problems relating to counterfeiting, and to the harsh treatment of those caught in the act, were, however, perennial.

In Scotland, by contrast, no note issuing monopoly existed, and banks were free to issue notes, although two banks dominated, namely, the Bank of Scotland and the Royal Bank of Scotland. Furthermore, several private note-issuing banks were in business in Ireland, and the Bank of Ireland was established in 1783. These banks relied on the Bank of England to obtain silver and gold, particularly during times of financial stress, such as 1783 and 1793.

Following a dramatic rise in government expenditures after 1793 due to the war against France, which caused a large rise in the Bank's note issue, the Bank's gold holdings fell sharply. After a scare about a French invasion convertibility was suspended in 1797, and resumed only in 1821. In view of the financial exigencies of the war, and the fact that there was in such circumstances no limit to the expansion of its note issue, now effectively legal tender, by the Bank, a privately owned company, what is in retrospect surprising about the period of suspension is how comparatively low the resulting inflation was. Even so, it was high enough to set off a major debate on its causation, for example in the Parliamentary Committee on the High Price of Bullion (1810). This period saw a further consolidation of the Bank as a note issuer, since it began to issue small denomination notes (given the shortage of silver and gold coin), which became legal tender in 1812. Furthermore, in 1816 silver coin ceased to be legal tender for small payments. The government also moved most of its accounts to the Bank in 1805 (in 1834 all government accounts were finally moved to the Bank).

During the 18th century and early part of the 19th century, smaller country banks had proliferated throughout England and Wales, many issuing their own notes. Given the prohibition on joint stock banking, the capital of these banks was usually small, and they regularly became insolvent, especially when the demand for cash (coin) became strong. This contrasted sharply with Scotland, where joint stock banking and branch banking were permitted, and relatively few failures occurred. Following a severe banking crisis in 1825, during which many English country banks failed, an Act renewing the Bank's charter (in 1826) abolished the restrictions on banking activity more than 65 miles outside of London. This led to the establishment of several joint stock banks, while the Bank countered by opening several branches throughout England.

Thus, a semblance of a banking 'system' began to emerge by 1830, with the Bank of England as the 'central' bank. By far the best book on such nascent central banking at this time was that written by Henry Thornton, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802). The practice of banks placing surplus funds with bill brokers also developed, with the Bank beginning to extend secured loans to these brokers on a more or less regular basis. In 1833 joint stock banks were finally allowed to operate in London, although they were not permitted to issue notes and thus were essentially deposit-taking banks only. The same Act specified that Bank of England notes were legal tender, and the Bank was also given the freedom to raise its discount rate freely (until then usury laws had placed a ceiling on interest rates) in response to cash outflows. The Bank's reaction (an early reaction function), in varying its interest rate, to cash inflows and outflows became codified around this time in what became known as the Palmer rule, after Horsley Palmer, Governor 1830–33, though the rule itself is usually dated from 1827.

The position of Bank of England notes was consolidated in an important Act, passed in 1844, generally known as the Bank Charter Act, preventing all note issuers from expanding their note issue above existing levels, and prohibiting the establishment of any new note-issuing banks. The 1844 Act also separated the issue and banking functions of the Bank into different departments, and required the Bank to publish a weekly summary of accounts.

Given that it did not pay interest on its deposits, the deposit activity of the Bank could never really

compete with that of other banks, which expanded rapidly from 1850 onwards. In 1854, joint stock banks in London joined the London Clearing House, and it was agreed that clearing by transfer of Bank of England notes would be abandoned in favour of cheques drawn on bank accounts held at the Bank. Ten years later the Bank of England itself entered this clearing arrangement, and cheques drawn on bankers' accounts at the Bank became considered as paid.

Although the Bank had, from the beginning of the 19th century, periodically bought or sold exchequer bills to influence the note circulation, explicit open-market borrowing operations to support its discount rate began in 1847. From 1873 until 1890 the Bank almost always acted as a borrower rather than a lender of funds, as there were typically cash surpluses. As a result, the Bank introduced the systematic issue of Treasury bills via a regular tender offer in 1877. Treasury bills had a much shorter maturity (three to twelve months) than Exchequer bills (five or more years), and were to play an important role in raising funds from the outset of the First World War onwards.

By 1890, the Bank's role as lender of last resort became undisputed when it orchestrated the rescue of Baring Brothers and Co., a bank whose solvency had become suspect, threatening to cause systemic problems. Earlier, in 1866, the failure of a discount house, Overend, Gurney and Co., had precipitated a financial panic, during which the Bank discounted large amounts of bills and extended considerable loans. The Bank, however, was criticized for not doing more to prevent the onset of such a panic, not least by Walter Bagehot in his famous book *Lombard Street* (1873).

Throughout the 19th century, the Bank streamlined its discount facilities. In 1851 it overhauled its discount rules, stipulating that only those parties having a discount account could present bills, and that these bills had to have a maturity of fewer than 95 days and be endorsed by two creditworthy firms. In the latter part of the century, however, the Bank gradually came to favour discount houses, often by presenting them with better rates of discount, and the range of firms doing discount business with the Bank declined. Discount houses were favoured because there was tension then between the Bank and the rapidly growing commercial banks – there was much banking consolidation via mergers between the 1870s and 1914 – and dealing via the intermediation of the discount houses enabled the Bank to influence market rates without having to interact directly with the joint-stock banks as counterparties.

Until the First World War the Bank pursued a discount policy which was primarily aimed at maintaining its gold reserves (as noted earlier) and which was conducted largely independently of the government. During the First World War, however, a clash occurred between the Bank Governor (Cunliffe) and the Chancellor (Law), during which the government made clear that it bore the ultimate responsibility for monetary policy, and that the Bank was expected to act on its direction.

A Subservient Bank, 1914–1992

The First World War was a major watershed not only in the history of the Bank but in the world more widely. It ushered in a half-century of increasing government intervention in every country, of a move towards socialist economies in most, and of communism in a wide swathe of countries. Under these circumstances the Bank became increasingly subservient to the government, in practice to the Chancellor of the Exchequer and to the Treasury, in the conduct of macromonetary policy, its previous primary function.

Initially, however, there was little perception that the war and the rise of socialist ideas had irretrievably altered the context for policy. There was a desire to return to the previous regime, the gold standard, with its tried and true verities, as expressed in the Cunliffe Committee Report (the first report of the Committee on Currency and Foreign Exchange 1919). That was probably inevitable under the circumstances, but a much more questionable decision was to return at the pre-war parity (against gold) despite the war-induced loss of markets (especially for the UK's main staples, textiles, coal, and iron and steel) and of competitiveness. Several of the other belligerent states,

notably France, had inflated, and allowed their exchange to float downwards by so much that they did not seek to re-peg at the previous parity, but could choose a more suitable and competitive rate. While the decision to return to gold at the pre-war parity, steadfastly supported by the Bank, has been much criticized, the modern theory of time inconsistency provides some defence, namely, if the Bank had started to change the chosen rate to suit the immediate conjuncture it would have been expected to do so again in future, making commitment to the regime less credible.

Be that as it may, conditions after the First World War, with a weak balance of payments and a massively inflated money stock and floating debt, were hardly conducive to the re-establishment of gold standard conditions. Indeed, the authorities initially felt forced to move in the other direction, to unpeg the sterling–dollar rate that had been established since 1916 and formally to leave the gold standard in March 1919. The ending of the war led then to an extremely sharp and short boom and bust, in which tight monetary policy played a major role in the subsequent deflation (see Howson 1975). From then until the return to gold at the pre-war parity of $4.86 to the pound in 1925, the Bank advocated keeping the Bank rate high enough to facilitate that regime change, but decisions on Bank rate and on the conduct of monetary policy were joint, in that no proposal by the Bank could be activated without the agreement of the Chancellor and HM Treasury; the Treasury view, however, then was in line with classical thought, namely, that monetary policy could and should impinge primarily on nominal prices, with real output affected by real factors.

Despite the boom in the USA, growth in the UK was perceived as remaining low and unemployment high, at least as compared with its main comparator countries, in the 1920s. This was in part due to the continuing problems of restoring a successful economic regime in Europe, wherein German reparations had a malign effect. Although the Bank had lost much of its power to direct domestic monetary policy (to Whitehall), the Bank and its Governor, Montagu Norman, played a leading role in the various international exercises to try to restore Europe to normality and to the gold standard, (Sayers 1976, ch. 8); and Sir Otto Niemeyer, a top Bank official, spread the gospel of establishing central banks to maintain price stability to the Dominions.

This whole structure came apart in the crisis that started in the USA in 1929 and then engulfed the rest of the world progressively through the subsequent four years. How far that collapse was itself exacerbated by the attempt to restore the gold standard has been explored by Eichengreen (1992). The UK was not in a strong economic position to avoid the world recession, but suffered a much smaller decline in output than in the USA or much of Continental Europe. The struggle to maintain the gold standard had required the maintenance of high interest rates, despite the imposition of controls on new issues in sterling by foreign governments. Despite high unemployment, wages and prices remained too sticky to allow the restoration of international competitiveness, though quite why this was so remains a debated issue.

With the gold standard collapsing in Europe and social pressures rising in the UK, there was diminishing political will to take the measures that appeared necessary to maintain the gold standard. The government decided to abandon it (in Norman's absence) in September 1931. From that moment onwards, until May 1997, the decision to alter the Bank rate moved decisively to Whitehall, effectively into the hands of the Chancellor, advised by HM Treasury. Of course, the Bank could, and did, make suggestions and played a major role in all the discussions, but the Chancellor took the decisions. Indeed, from June 1932 until November 1951 a policy of cheap money was followed whereby Bank rate was held constant at two per cent. Norman stated in 1937, 'I am an instrument of the Treasury'.

Meanwhile, the Bank was becoming more professional. The old system of circulating the Governor's chair in turn among the directors of the Bank, who were appointed from city (but not commercial bank) institutions, was superseded by the continuing governorship of Montagu Norman from 1920 until 1944. While this arose by happenstance rather than intention (see Sayers 1976, ch. 22), it gave the Bank highly skilled,

even if also highly idiosyncratic, leadership. Moreover, Norman introduced economists and other able officials into both the staff and the Court (the largely ceremonial board) of the Bank, although it is (apocryphally) recorded that Norman told one such economist, 'You are not here to tell me what to do, but to explain why I have done what I have already decided to do.'

In effect, the Bank had already become nationalized by the end of the Second World War. So the formal act of nationalization in 1946 brought about no real substantive changes, except that the Governor and his deputy (there has as yet been no woman Governor, although Rachel Lomax became the first female Deputy Governor in 2003), were appointed by the government for five years, renewable once more in most cases. Indeed, the more profound changes were brought about by Governor Gordon Richardson (1973–83) in the early 1980s. Until then, the Governor had been rather akin to a chairman, with the deputy and other internal directors as members of the board, setting strategy. Much of the executive power still lay with the Chief Cashier, who acted as leader of the heads of department, who ran the Bank. There was a clear break, a division, between the staff in the departments on the one hand and the Governors and Directors on the other. Richardson changed all that, concentrating power in the Governors' hands, sharply demoting the role of Chief Cashier, and underlining the precedence of (internal) directors over heads of department in all policy matters.

So, as power to decide the course of monetary policy – and to set the Bank rate passed to Whitehall, what did these professional central bank officials do? The Bank came to have three main areas of responsibility. The first was the management of markets, notably the money market, the bond (gilts) market and the foreign exchange market. The UK had come out of the Second World War with a massively inflated ratio of debt to GDP, and its management had remained difficult and delicate, at least until after the War Loan Conversion of 1932. No sooner, however, had debt management been thereby put on a sounder foundation than the Second World War led to a further upsurge in the debt ratio, which led once again to debt management becoming a major preoccupation of policy. Thereafter, a combination of generally prudent fiscal policies, so that the debt ratio fell steadily, and then unexpected inflation in the 1970s, which accelerated the decline in the debt ratio, and market reforms in the 1980s, enabled the procedures of debt management to become simpler and standardized. Similarly, the floating exchange rate in the 1930s, followed by attempts to maintain pegged exchange rates both during the Second World War and thereafter under the Bretton Woods system, against a background of perennially weak balance of payments conditions, made the management of the UK's foreign exchange reserves and intervention on the foreign exchange market a crucial function of the Bank until 1992, when the UK was forced out of the European exchange rate mechanism. During crises the officials in charge of such foreign exchange operations were in telephone communication with the Chancellor and, occasionally, the Prime Minister at frequent intervals.

The Bank held that such market operations required a special professional expertise (though HM Treasury remained sceptical). The Bank threw itself into such activities with enthusiasm, and defended its pre-eminent role in this respect stoutly against all outside encroachment or criticism. Indeed, its market 'savvy' was its most powerful lever to persuade the Chancellor to its views in any debate; 'I am sorry, Chancellor, but the market will not accept that policy' was the strongest card it had to play, and that card was played often and with alacrity.

Although ultra-cheap money, with Bank rate held at two per cent, was abandoned in 1951, when the Conservative Party was returned to office, monetary policy in general, and interest rates in particular, were still seen as both more ineffective and uncertain in their impact on domestic demand than the supposedly more reliable fiscal policy, a conclusion upheld by the controversial Radcliffe Report (1959). Consequently, fiscal policy was used to try to steer domestic demand while interest rates were raised to protect the balance of payments during the regular bouts of external weakness, and otherwise held low both to ease government finance and to

support fixed investment. The outcome was a system in which inflationary pressures regularly threatened both the internal and external value of the currency. The chosen solution was to supplement market measures by direct interventions, in the case of external pressure via exchange controls, in the case of monetary expansion via direct controls on bank lending to the private sector. In both instances the Bank acted as the administrative agent of HM Treasury.

Such direct controls were introduced (on bank lending), or greatly extended and tightened (exchange controls), with the onset of the Second World War in 1939, but were continued, for the reasons outlined above, until 1971 for bank lending and 1979 for exchange controls. The administration of exchange controls required a large staff, but, unlike with its market operations, the Bank had little enthusiasm for acting in this guise. The Bank hoped to restore London to its former role as an international financial centre. While it succeeded in this through its encouragement of the Eurodollar market, aided by inept US policies, the continued administration of exchange controls remained an unwelcome burden. The same was true for direct controls on bank lending. Such controls were regarded by politicians as a comparatively painless way of dampening demand and inflation, while they were resented by commercial bankers. The Bank found itself in the middle of these disputes, and grew painfully aware of such controls' stultifying effect on efficiency, dynamism and growth. The Bank, inspired by John Fforde (the then executive director in charge of domestic finance, and subsequent Bank historian), pressed hard for these controls to be dismantled, and succeeded with the liberalizing reform of Competition and Credit Control (Bank of England 1971).

As with many other cases of banking liberalization, such as in Scandinavia at the end of the 1980s, this was followed by an expansionary boom and then a bust, the fringe (secondary) bank crisis of 1973/74 (Reid 1982). While there remain questions about how monetary policy could have been better applied to prevent the prior monetary boom (1972/73), there was no question but that the financial crisis found both the Bank and the banks unskilled in risk management and unprepared for adverse shocks to financial stability. The long period of financial repression – that is, controls on bank lending to the private sector and force-feeding with government debt – had had the by-product of making the (core) commercial banking system safe between the mid-1930s and the early 1970s. The central banking function of maintaining financial stability, via regulation and supervision, had atrophied.

This had not been so earlier, and the Bank had been closely involved in the rescue of Williams Deacon's Bank by the Royal Bank of Scotland in 1930 (Sayers 1976, ch. 10), and in helping to shape the structure of both the commercial banking system and the London Discount Market Association. Williams Deacon's had got into trouble largely because of bad debts from Lancashire cotton companies. Norman, and the Bank, extended their structural interventions beyond banking to try to encourage strategic amalgamations to shore up the positions of weakened companies in a variety of industries, such as cotton, steel, shipping, armaments (Sayers 1976, ch. 14). The Bank's involvement in structural matters outside of banking itself was episodic depending on both circumstances and personalities. Another example of such Bank involvement was the considerable role it played in the reform of the UK capital market in the 1980s, more familiarly known as 'Big-Bang'. But views on whether the Bank has any locus in such wider structural issues vary over time; the early 2000s saw a major withdrawal by the Bank from any such involvement.

The fringe bank crisis in the early 1970s was, however, a clarion call to put more emphasis on its third main function, bank supervision and regulation. The immediate result was a reorganization in the Bank. Initially a nucleus of a new specialized department was established in the Discount Office where the limited staff assigned to this role had sat, which rapidly absorbed staff and resources. Thereafter this became a separate department devoted to banking supervision and regulation (its first head was George Blunden, later to become Deputy Governor, who handed it on to Peter Cooke in 1976). Its position was regularized in the Banking Act (1979) which gave formal

powers to the Bank to authorize, monitor, supervise, control and, under certain circumstances, withdraw prior authorization (tantamount to closure) for banks. No such powers had been available before that date. Meanwhile, other financial intermediaries, such as building societies or insurance companies, remained (lightly) regulated by various government departments.

The fringe bank crisis was almost entirely domestic, confined to British headquartered companies. Meanwhile, however, the onwards march of liberalization (involving the removal of direct controls, notably exchange controls in 1979) and of information technology were leading to a growing internationalization of financial business. For a variety of reasons, mostly relating to the innovation of the Eurodollar and Euro-markets, London regained its role as an international financial centre in the 1960s, and thus international monetary problems became of particular importance to the Bank, which took a leading role in such matters from the 1970s onwards.

Central bankers had met regularly at the headquarters of the Bank for International Settlements (BIS) in Basel for many years. It was, therefore, a logical step for supervisory officials also to come together at Basel on regular occasions to discuss matters of common interest. Thus was born (in 1974), as a result of an initiative from Gordon Richardson, the Basel Committee on Banking Regulation and Supervisory Practices. For the first 15 years of its existence it was chaired by the participant from the Bank of England, and was usually known by his name; thus, the Blunden Committee (1974–77) gave way in due course to the Cooke Committee (1977–88). The failures of Franklin National and Herstatt prompted the First Basel Concordat, which allocated responsibility for supervising internationally active banks to home and host authorities.

So by the mid-1970s, a need was perceived for banking supervision at both the domestic and, via consolidation, at the international levels. The purpose of these initiatives was to clarify where responsibility lay for the supervision of international banks, to prevent fragile, and possibly fraudulent, banking leading to avoidable failures and potential systemic crises.

Despite the growing number of bank supervisors, and notable success in reversing prior declines in capital ratios, the history of banking in the subsequent decades in the UK was spotted by occasional bank failures. Unlike the fringe bank crisis, none was, or was allowed to become, systemic, nor did individual depositors lose any money, except in the case of Bank of Credit and Commerce International (BCCI), and even in that case the deposit protection scheme provided some relief. The failures of Johnson–Matthey (in 1984), BCCI (in 1991) and Barings (in 1995) were all isolated cases of bad, in some respects fraudulent, banking.

The main problem of the 1970s and 1980s was, however, that of combating inflation, which soared to heights previously unknown, not only in peacetime but even in wartime, during the 1970s, up to 25 per cent per annum. There were three main theories, though divisions between them were never completely distinct. The first was the cost-push theory, that inflation was driven by over-mighty trade unions, seeking to increase the relative real pay of their members; the appropriate remedy was then prices and incomes policies plus reform (and constraint) of trades unions. The second was the (vertical) Phillips curve analysis; the remedy here was to raise unemployment above the 'natural' rate to reduce inflation. The third was that inflation was a monetary phenomenon; the remedy was to control the rate of growth of the (appropriate) monetary aggregate.

Until the mid-1970s, both major political parties, the Bank and HM Treasury all professed some combination of theories 1 (cost-push) and 2 (Phillips curve). Left-leaning politicians, academics and officials tended to put more weight on cost-push. In the 1960 and 1970s the third, monetarist, view seemed to explain events better and gained strength, not only in the USA (Milton Friedman) but also in the UK. In particular, the surge in inflation in the UK in 1973–75 followed closely behind the rapid expansion of broad (but not narrow) money in 1972–73. So, when in opposition, the leading Conservative politicians Keith Joseph and Margaret Thatcher embraced a version of monetarism.

When they came to power in 1979, they tried to commit monetary policy to follow a target for broad money, via the Medium Term Financial Strategy. In order to achieve this, nominal, and real, interest rates were kept high, and the exchange rate appreciated sharply, partly under the influence of North Sea oil and confidence in Thatcherite policies. Inflation duly declined, as planned, but broad money growth did not. This latter was partly due to the abolition of the 'corset' in 1980. The 'corset' was a reformulated, and somewhat disguised, direct control over commercial bank expansion that had been pressed into service on several occasions during the 1970s. The Bank was glad to see the end of exchange controls and direct controls over bank lending, but had never shared the government's monetarist faith in trying to set, and stick to, targets for the growth of (the various) monetary aggregates.

The empirical demonstration of the unpredictability of the relationship between (broad) money and nominal incomes in the early 1980s soon weakened the government's own faith. After moving from one monetary target to several joint targets, and an attempt to hit the broad money target by 'overfunding', an exercise criticized by many as artificial, the government abandoned its monetary targetry in 1986.

That left the question of how monetary policy, and with it control of inflation, was to be managed or, in the standard phrase, 'anchored'. The then Chancellor, Nigel Lawson, wanted to 'anchor' by joining the exchange rate mechanism (ERM) of the European Monetary System and leaving the steering of monetary policy to the Bundesbank. The Prime Minister, Mrs Thatcher, and her adviser, Alan Walters, were opposed, both on economic grounds (that such a pegged system was 'half-baked') and for wider political reasons. There was a battle royal in which the Bank was left on the sidelines. Lawson was sacked, but eventually Mrs Thatcher was, grudgingly, persuaded to allow the UK to join the ERM in October 1990.

This was in the aftermath of German reunification, and the expenditures connected with that led the Bundesbank to keep interest rates higher than was tolerable for the UK (or Italy). The UK was in the throes of a sharp downturn in housing prices, following an unstable housing boom in the late 1980s. With the Conservatives having become politically weaker, there was just no stomach to raise interest rates to the levels necessary to sustain the ERM. The UK was forced out in September 1992.

Independent and Focused, 1992–

The ejection of the UK from the ERM left the government and HM Treasury with the recurrent problem of how to manage, to 'anchor', monetary policy. Both monetary and exchange rate targets had been tried, and both had been found wanting. While the economic experience of the 1980s was better than that of the stagflationary 1970s, it was hardly stellar, with a boom–bust cycle at the end of the decade.

Meanwhile, a new approach had been adopted in New Zealand, whereby the central bank was given administrative freedom to vary interest rates for the purpose of hitting a target for the inflation rate, jointly set by the government and the central bank: that is, inflation targetry. This obviated one of the shortcomings of monetary targetry, namely, the unpredictability of the velocity of money; it left setting the goals of policy, the overall strategy, in the hands of government, but shifted the (constrained) discretion to vary interest rates to the professional and technical judgement of the central bank. This procedure soon generated a strong body of academic support (for example, Fischer 1994).

Although Conservative Chancellors (both Lawson and Lamont) had toyed with the idea of giving the Bank operational independence, consecutive Prime Ministers (Thatcher and Major) refused, primarily on political grounds. Nevertheless Lamont wanted to move to an inflation target. But there was a problem of governmental credibility. To foster credibility, Lamont now encouraged (in 1992/93) the Bank to prepare and to publish an independent forecast of the likely projection for inflation, the *Inflation Report* (on the assumption of unchanged policies); this was a reversal of prior habits whereby HM Treasury

and Ministers customarily censored Bank publications and discouraged any publication of internal Bank forecasts. The process of gradually giving the Bank a more independent role in setting monetary policy took a step further when the next Chancellor, Clarke, not only held a meeting with the Governor, and the Bank, to discuss future changes in interest rates, but published the minutes of the meeting, including the Governor's initial statement, verbatim; this was termed the Ken (Clarke) and Eddie (George) show. That said, Clarke had strong views on the appropriate policy and on a couple of occasions overruled the Governor's suggestions.

At that time – the mid-1990s – there were still question marks over the Labour Party's ability to manage the economy; financial markets are inherently suspicious of left-leaning governments. So Labour had more to gain (than the Conservatives), in terms of confidence and lower interest rates, by granting operational independence (back) to the Bank. In advance of the 1997 election the then shadow Chancellor, Gordon Brown, was cautious; while indicating general support for both inflation targetry and operational independence, he stated that he wanted time to see how well the Bank performed before granting such independence. But, within days of winning the election, he made that strategic change to the monetary regime.

This was, of course, a great prize for the Bank, but it did not come without cost. In the same month as operational independence was awarded to the Bank, both debt management and banking supervision were hived off, to a separate Debt Management Office (DMO) and Financial Services Authority (FSA) respectively. With the government debt to GDP ratio having declined and capital markets strengthened, debt management had become more of a routine and standardized exercise. Nevertheless, its departure to the DMO, and the fact that the float of the exchange rate after 1992 was kept 'clean', that is, without intervention, meant that much of the market operations which had been so central to the Bank in the post-Second World War period disappeared, though its money market operations, of course, continued. The administration of direct controls had gone at the beginning of the 1980s. And now banking supervision was also taken away. This meant that almost *all* the prime functions that the Bank had undertaken in its post-Second World War period of subservience had now gone. Instead, the Bank was now focused on varying interest rates to achieve the inflation target set for it by the Chancellor.

There are numerous arguments, quite evenly balanced, for whether bank supervision should be kept within a central bank or put with a separate Financial Services Authority (FSA), covering both banks and other financial intermediaries (see Goodhart 2000). Be that as it may, there are various aspects of the financial system, such as oversight of the payments' system, and of crisis management, such as lender of last resort functions, which cannot be delegated to an FSA. Moreover, the achievement of price stability is likely to be seriously compromised by any serious bout of financial instability – and vice versa, with financial stability adversely affected by price instability. So the removal of individual bank supervision does not absolve the Bank from concern with financial stability issues more widely; indeed, the Bank is specifically charged with maintaining overall systemic stability in the financial system. But exactly what that means when responsibility for the conduct of individual bank supervision is located elsewhere is not yet entirely clear.

What it certainly does mean is that the FSA, the Bank, and the political authorities as the ultimate source of any needed fiscal support have to work extremely closely together, in advising on any new regulations (whether domestic or international), in monitoring developments (as in the Financial Stability Review), and in crisis management. This latter task would be done via the Tripartite Standing Committee (FSA, Bank, and HM Treasury), set up in 1997, although so far no such financial (as contrasted with simulated 'war games') crisis has occurred, though the Committee did meet after the terrorist attacks on 7 July, 2005. How successful crisis management by such a committee may be has yet to be seen.

The monetary policy function of the Bank, now its central preoccupation, has, however, been very successful by all the usual criteria. In several papers Luca Benati (for example, Benati 2005) has demonstrated that the variance of both GDP and of inflation around its target has been lower under the inflation targetry regime (whether taken as starting in 1992 or in 1997) than under any previous historical regime. The procedures of having a Monetary Policy Committee consisting of five senior Bank officials and four outside experts (appointed by the Chancellor), with the Committee serviced by Bank staff, has worked generally smoothly and well. So the Bank's reputation and credibility have rarely been higher, although now tightly focused on one main function.

See Also

- ▶ Banking Crises
- ▶ Bullionist Controversies (Empirical Evidence)
- ▶ Gold Standard
- ▶ Inflation Targeting
- ▶ Monetary Policy, History of

Bibliography

Acres, W. 1931. *The Bank of England from within*. London: Oxford University Press.

Andréadès, A. 1909. *A history of the Bank of England*. London: P. S. King and Sons.

Bagehot, W. 1873. *Lombard street*. London: Kegan, Paul and Co.

Bank for International Settlements. 1963. Bank of England. In *Eight European central banks*. Basle: Bank for International Settlements.

Bank of England. 1971. *Competition and credit control*. London: Bank of England.

Benati, L. 2005. The inflation-targeting framework from an historical perspective. *Bank of England Quarterly Bulletin* 45(2): 160–168.

Bowman, W. 1937. *The story of the Bank of England: From its foundation in 1694 until the present day*. London: Herbert Jenkins.

Chapham, R. 1968. *Decision making: A case study of the decision to raise the bank rate in September 1957*. London: Routledge and Kegan Paul.

Clapham, J. 1944. *The Bank of England: A history*. Cambridge: Cambridge University Press.

Clay, H. 1957. *Lord Norman*. London: Macmillan.

Committee on Currency and Foreign Exchange After the War (Cunliffe Committee). 1918. *First Interim Report*, Cmnd. 9182; and 1919. *Final Report*, Cmnd 464. London: HMSO.

Eichengreen, B. 1992. *Golden fetters: The gold standard and the great depression*. New York: Oxford University Press.

Feavearyear, A. 1963. *The pound sterling: A history of English money*, 2nd edn, rev. E. Morgan. Oxford: Clarendon.

Fforde, J. 1992. *The Bank of England and public policy 1941–1958*. Cambridge: Cambridge University Press.

Fischer, S. 1994. Modern central banking. In *The future of central banking*, ed. F. Capie, C. Goodhart, S. Fischer and N. Schnadt. Cambridge: Cambridge University Press.

Geddes, P. 1987. *Inside the Bank of England*. London: Boxtree.

Giuseppi, J. 1966. *The Bank of England: A history from its foundation in 1694*. London: Evans Brothers Limited.

Goodhart, C. 2000. *The organisational structure of banking supervision,* Special paper, no. 127. London: Financial Markets Group Research Centre, London School of Economics. Subsequently published in *Economic Notes* 31: 1–32.

Hennessey, E. 1992. *A domestic history of the Bank of England 1930–1960*. Cambridge: Cambridge University Press.

Howson, S. 1975. *Domestic monetary management in Britain, 1919–38*. Cambridge: Cambridge University Press.

Radcliffe Report. 1959. *Report: Committee on the working of the monetary system*, Cmnd 827. London: HMSO.

Reid, M. 1982. *The secondary banking crisis, 1973–75: Its causes and course*. London: Macmillan.

Richards, R. 1929. *The early history of banking in England*. London: Frank Cass and Co.

Rogers, J. 1887. *The first nine years of the Bank of England*. Oxford: Clarendon.

Sayers, R. 1936. *Bank of England operations, 1890–1914*. London: P.S. King and Son.

Sayers, R. 1957. *Central banking after Bagehot*. Oxford: Clarendon.

Sayers, R. 1976. *The Bank of England, 1891–1944*. Cambridge: Cambridge University Press.

Smith, V. 1936. *The rationale of central banking*. London: P.S. King and Son.

Steele, H., and F. Yerbury. 1930. *The old bank of England*. London: Ernest Benn.

Stockdale, E. 1967. *The Bank of England in 1934*. London: Eastern Press.

Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*. New York: Kelley, 1962.

Ziegler, D. 1990. *Central Bank, peripheral industry: The Bank of England in the provinces 1826–1913*. London: Leicester University Press.

Bank Rate

A. B. Cramp

This was the label applied to the rate at which the Bank of England would discount first-class bills of exchange in the London market: by extension, it has come to mean the rate at which any central bank makes short-term loans available to domestic commercial banks. The UK Bank Rate's practical significance dates from the Bank Charter Act of 1833, Section 7 of which exempted bills of a currency up to three months from the provisions of usury laws which had previously imposed a 5 per cent interest ceiling. This relaxation had been recommended in 1802 by Henry Thornton as a means of containing demand for discounts, which passed along a chain from country banks to London banks to the nascent last-resort central bank, and threatened to become excessive when market forces would have pushed rates above the ceiling. The urgency of such containment was increased as a result of (a) these 'internal' gold drains being reinforced by 'external' analogues related to the expansion of international trade and capital movements; (b) the imposition by the 1844 Bank Charter Act of a limit to the fiduciary issue, of Bank of England notes backed by holdings of securities, designed to ensure the maintenance of convertibility of notes into gold. The 1847 liquidity crisis forced the Government to promise a retrospective act of indemnity should this limit be breached, freeing the Bank to act as lender of last resort to whatever extent the exigencies of the crisis might require – but on condition that a Bank Rate of not less than 8 per cent be imposed.

Henceforward, and until the final abandonment of the gold standard in 1931, Bank Rate changes were the major technique by which the Bank of England protected its reserve. The technique was powerful at least until the First World War, after which its effectiveness was compromised by political and economic disorder, and by the rise of New York as an international financial centre alternative to London. Understanding of the causes of the pre-1914 power of Bank Rate increases (reductions tended to represent rather passive reactions to relaxation of pressures) is facilitated by distinguishing responses in the spheres of, respectively, the London money market; external trade and payments; and internal economic activity.

Within the London money market, matters hinged – in the manner adumbrated by Thornton – on bankers' response to the rise in Bank Rate to a 'penalty' level, above the market rate(s) at which the bankers and themselves acquired bills. Bank Rate thus operated, in Walter Bagehot's phrase, as a 'fine on unreasonable timidity' in regard to the liquidation of banks' assets with a view to strengthening reserve ratios, against the possibility of a run on banks by nervous depositors. Originally, it is to be noted, the initiative lay with the commercial banks rather than with the developing central bank; the shortage of cash (=deposits at the Bank of England) resulted from increased demand by the former, rather than from reduction of supply engineered by the latter; autonomous pressures were already raising (short-term) interest rates, and Bank Rate changes were an important – probably overriding – influence on the extent of the rise by virtue of the Bank of England's position as key supplier of an essential margin of funds. There was thus no real problem in 'making Bank Rate effective', that is to say ensuring that it exerted appropriate influence on market rates. Nor was there any call for assistance from the weapon, not in any case developed until after World War I, of open-market sales of securities at central bank initiative. These points warn modern theorists against the temptation to read back into the 19th century later-developed notions suggesting that the rise in *price* (short-term interest rates) either reflected, accompanied or caused a reduction in *quantity* (bank credit flows, or bank deposit totals). The relationship between Bank Rate changes and 'the quantity of money' was, as Keynes argued (see below) much more diffuse and complex than modern monetarist styles of theory can easily envisage; its character can hardly begin to emerge until repercussions outside the money market have been considered.

Of these repercussions, those relating to external flows, rather than to internal adaptations, were the main focus of attention in Bank Rate's classical period, and we first consider the external side. Ricardian thought, in the early part of the period, encouraged attention to the trade balance; but in practice, as the 19th century wore on, the action was increasingly seen to occur in the sphere of international payments and capital movements. This was mainly a reflection of structural changes which produced a consistently strong UK trade balance, massive long-term overseas lending, and a growing mass of internationally mobile bills of exchange (principally the 'bill on London'). It was also, by the turn of the century, a reflection of (probably fortuitously) helpful policy by the Bank of France, the focal point of London's only rival as a financial centre. The Bank of France kept more substantial gold reserves than the Bank of England; and it was willing to allow those reserves to vary in order to exert stabilizing influence on continental interest rates. As a result, a rise in London's Bank Rate tended to increase the differential between UK and foreign short-term rates, and to tilt the balance of short-term flows in London's favour. An increase in Bank Rate, opined the Cunliffe Committee in 1918, would 'draw gold from the moon'; in practice, the metal did not travel quite so far.

A highly significant implication of this (at the time, ill-understood) conjuncture, was that the Bank of England discovered a power to protect its reserve without significant damage to UK overseas trade. The validity of this judgement is witnessed by the decline in the volume of complaints from traders about the burden of high short-term interest rates. Such complaints were quite substantial in the early decades of intermittently high and rising Bank Rate levels. The present author has established (1962), however, that the grievances were much more closely related to the *availability* of short-term credit than to its *cost*. A rise in Bank Rate (from even quite low levels) was seen, with good reason, as heralding a potential liquidity shortage that might be transformed quickly into a liquidity crisis: alert bankers and traders at once began to exercise caution in undertaking new commitments. This is undoubtedly the historical origin of what would otherwise be a rather puzzling strand in the Bank Rate tradition, namely the idea that a rise in Bank Rate operated as an 'Index', a storm signal enjoining caution. This strand persisted in financiers' folk-memories long after its realistic institutional basis had declined, and resurfaced in the 1950s in a new form: sterling crises could be countered by a 'package deal' of measures, of which a Bank Rate rise constituted an essential element, as an *index* of the UK authorities' determination to inflict whatever pain might be necessary to rectify external imbalance.

In just what this pain might consist had been a matter of debate, intermittently vigorous, among academic economists – whose primary attention, in the 20th century, came to focus on the internal economy, and the effects thereon of what the 1918 Cunliffe Committee saw as a Bank Rate-induced (? accompanied) general rise of interest rates and restriction of credit. The emphasis on credit restriction was by then probably exaggerated, and traceable to the folk-memories just noted. The emphasis on generally rising interest rates undoubtedly exaggerated Bank Rate's *direct* influence on the structure of interest rates. It is true that, by 1900, commercial bank borrowing and lending rates were widely (not universally) linked to Bank Rate – an administrative link reflecting a market reality for, as indicated above and as Bagehot had argued, an institution (the Bank of England) that regularly supplied the market with the necessary residual margin of cash almost automatically exercised what we should call 'price leadership', its own price for short-term accommodation dominating other influences. Keynes was thus justified, in his *Treatise on Money* (1930), in treating Bank Rate as representative of the general level of *short* rates, on the assumption that Bank Rate changes were normally 'effective' in influencing market rates. The further link to *long* rates, however, was more problematic, and a source of disagreement between Keynes and R.G. Hawtrey (1938).

Hawtrey tended to downplay the link, on the argument that the direct influence on long rates of a rise in short rates depended on the period for which the rise was expected to last – which period,

because of Bank Rate's external power described above, was typically brief. His view was doubtless influenced by his tenacious, and fairly isolated, adherence to the theory that Bank Rate's external power was mediated primarily by its influence on the cost of holding inventories. His theory was that individual merchants would have a strong inducement to respond to a Bank Rate increase by reducing purchases from manufacturers, designed to effect a temporary reduction of inventory levels during the limited period for which the higher Bank Rate was expected to last. But collectively these mercantile responses so reduced demand that manufacturers restricted their purchases of raw materials from merchants, and the 'vicious circle of deflation' was joined. Hawtrey claimed support for his theory from oral testimony, notably before House of Commons committees of inquiry into liquidity crises. But later investigation (Cramp, 1962) demonstrated that John Torr, Chairman of the Liverpool Chamber of Commerce during the 1857 crisis, was typical in arguing that what mattered to traders was 'not so much the rate of interest as the impossibility of getting the medium of exchange', that is, not so much the cost of credit as its availability, which gradually became more reliable as the techniques of commercial and central banking improved.

It was Keynes's view, in the *Treatise* and in the Report of the Macmillan Committee which he dominated, that exercised the more substantial and enduring influence on academic opinion. Unlike Hawtrey, he tended to emphasize the link through to long-term interest rates, perhaps implicitly assuming – by this juncture – the support of appropriate open-market operations, security sales by the central bank. He was by this stage urging that such sales should include bonds as well as bills, facilitating direct influence on long rates. Such advocacy was not uncongenial to a central bank now ever-anxious to 'fund the floating debt', reflecting fears of repetition of the experience of feeling constrained by government borrowing needs during the inflationary boom of 1920–21.

Keynes was thus enabled to presume that a rise in Bank Rate would be accompanied by supporting measures appropriate to the exertion of a strong *indirect* influence on the structure of interest rates. In this way, he justified retrospectively the Cunliffe Committee's rejection of Alfred Marshall's dismissal of the effect of Bank Rate changes as 'a ripple on the surface', and also inaugurated the era of academic preoccupation with the link between 'the rate of interest' (essentially, the long-term rate) and the level of expenditures on fixed investment. He contended (*Treatise*, I, pp. 154–5) that 'a rise in Bank rate tends, in so far as it modifies the effective rates of interest, to depress price levels'.

The theoretical model deployed to explain this proposition is significant for the history of monetary theory as well as that of Bank Rate. Keynes appealed to Wicksell's celebrated (1898) concepts, to argue that a Bank Rate increase represented a rise in the market rate of interest, relative to the natural rate which would equate desired levels of investment and saving. The link to prices, however, would come principally, not through the monetary route of reduced banklending flows and bank-deposit stocks, but through the impact of higher market interest rates on the decision to invest. A higher rate of discount would be applied to the stream of future yields anticipated from an act of investment. Such acts would be postponed, the more readily when the higher Bank Rate was regarded as a temporary divergence from the normal level, the more ineluctably on account of the likely difficulty in such market conditions of floating new issues on the capital market. Aggregate demand and prices would thus tend to be depressed, by processes which would result in reduced demand for money balances. The money market tightness would be superficially eased from the domestic side, as it would also be relieved from the foreign side – quickly on account of reduced lending to overseas borrowers, more slowly and fundamentally as the domestic deflation improved the trade balance.

The *General Theory*, of course, was soon to initiate a prolonged phase of even greater scepticism about the strength of the linkage between money and prices. It appeared at a time when cheap money was also causing de-emphasis on the role of changes in Bank Rate. From 1932 to

1951, Bank Rate was held, apart from a hiccough when war began in 1939, at the level of 2 per cent. Academic discussion continued of the relationship between the level of interest rates and decisions to invest, but it was largely severed from consideration of money-market techniques and policies. When inflationary fears began to surface late in the cheap money era, as Professor R.S. Sayers (1979) notes, D.H. Robertson 'addressed the world not on the question "What has happened to Bank Rate?" but "What has happened to the Rate of Interest?"'

The desire to restrain inflationary tendencies prompted the beginning in 1951 of a period of experimentation with the revival of monetary policy techniques, a trend which within a decade or so was to receive very substantial impetus from the anti-Keynesian monetarist counter-revolution originating principally in Chicago. In the earlier phases of this postwar period, Bank Rate changes were reintroduced to the authorities' armoury of measures, but somewhat tardily and half-heartedly, being subordinated to the then still quite fashionable preference for direct controls, e.g. on the volume of bank advances. As noted above, there was some disposition to regard a Bank Rate increase as an essential element in a restrictive 'package deal', but no-one seemed quite sure why, except that folk-memories even yet favoured it (*those* were the days, when even gold on the moon was magnetized!), and market enthusiasts instinctively welcomed a price element in a package consisting primarily of quantity controls. In the later, monetaristinfluenced, phases of the postwar period, quantity controls were precisely what influential opinion desired, but because that opinion favoured achieving them by market rather than by administrative measures, interest-rate changes were acknowledged to have a significant, though subsidiary, role.

Thus was Keynes's sequence, which as we have seen began from Bank Rate, reversed. Bank Rate was renamed, under the 'Competition and Credit Control' regime operated in the UK in the 1970s. It became 'Minimum Lending Rate' (MLR). It was ostensibly linked to the Treasury Bill rate emerging from the weekly tender, and consequently moved much more frequently than of yore, although every so often the authorities uncoupled the link, when they desired an old-fashioned 'index effect' – on external fund flows – from a rise in short-term rates clearly engineered by themselves.

Under the new (and nameless) UK monetary control régime of the 1980s, the ghost of Bank Rate became yet more evanescent. The continuous posting of MLR was formally suspended, though the authorities reserved the right 'in some circumstances to announce in advance the minimum rate which, for a short period ahead, it would apply in lending to the market'. This right has on occasion been actified. Bank Rate lives, just. Treatises on money no longer contain, as did Keynes's, a chapter on its *modus operandi*. But as in so many directions in economics, it would be a bold observer who projected the existing trend indefinitely, and predicted Bank Rate's final demise. There are continuities in economics, albeit disguised by irregular cycles in opinion and practice; trends persist, even in a new high-technological age.

See Also

- ▶ Cheap Money
- ▶ Dear Money
- ▶ Monetary Policy

Bibliography

Bank of England. 1971. Competition and credit control. *Quarterly Bulletin*.
Cramp, A.B.. 1962. *Opinion on bank rate 1822–60*. London: G. Bell.
Cunliffe (Lord), et al. 1918. *Committee on currency and foreign exchanges, first interim report*. London: HMSO.
Hawtrey, R.G. 1938. *A century of bank rate*. London: Longman.
Keynes, J.M. 1930. *A treatise on money*. London: Macmillan.
Keynes, J.M. 1936. *General theory of employment, interest and money*. London: Macmillan.
Sayers, R.S. 1981. *Bank rate in Keynes's century*. London: The British Academy.
Wicksell, K. 1898. *Interest and Price*. Trans. R.F. Kahn, London: Macmillan for the Royal Economic Society, 1936.

Banking Crises

Charles W. Calomiris

Abstract

Banking crises take a variety of forms ranging from temporary liquidity crises to massive insolvencies. They sometimes coincide with other financial crises in currency and sovereign debt markets, and sometimes occur in isolation. These differences reflect the variety of causal influences that give rise to problems for banks. The unusually crisis-prone experience of the United States historically reflected its unique industrial organization of banking. Policies intended to reduce the incidence of banking crises (especially deposit insurance) have instead often increased the risk of crises, as safety-net protection reduces market discipline, allowing banks to undertake imprudent risks.

Keywords

Banking crises; Central banks; Currency crises; Deposit insurance; Devaluation; Federal Reserve System; Great depression; Liquidity crises; Panic of 1907; Prudential bank regulation; Sovereign debt

JEL Classifications

N2

There are two distinct phenomena associated with banking system distress: exogenous shocks that produce insolvency, and depositor withdrawals during 'panics'. These two contributors to distress often do not coincide. For example, in the rural United States during the 1920s many banks failed, often with high losses to depositors, but those failures were not associated with systemic panics. In 1907, the United States experienced a systemic panic, originating in New York. Although some banks failed in 1907, failures and depositor losses were not much higher than in normal times. As the crisis worsened, banks suspended convertibility until uncertainty about the incidence of the shock had been resolved.

The central differences between these two episodes relate to the commonality of information regarding the shocks producing loan losses. In the 1920s, the shocks were loan losses in agricultural banks, geographically isolated and fairly transparent. Banks failed without resulting in system-wide concerns. During 1907, the ultimate losses for New York banks were small, but the incidence was unclear *ex ante* (loan losses reflected complex connections to securities market transactions, with uncertain consequences for some New York banks). This confusion hit the financial system at a time of low liquidity, reflecting prior unrelated disturbances in the balance of payments (Bruner and Carr 2007).

Sometimes, large loan losses, and confusion regarding their incidence, occurred together. In Chicago in mid-1932 losses resulted in many failures and also in widespread withdrawals from banks that did not ultimately fail. Research has shown that the banks that failed were exogenously insolvent; solvent Chicago banks experiencing withdrawals did not fail. In other episodes, however, bank failures may reflect illiquidity resulting from runs, rather than exogenous insolvency.

Banking crises can differ according to whether they coincide with other financial events. Banking crises coinciding with currency collapse are called 'twin' crises (as in Argentina in 1890 and 2001, Mexico in 1995, and Thailand, Indonesia and Korea in 1997). A twin crisis can reflect two different chains of causation: an expected devaluation may encourage deposit withdrawal to convert to hard currency before devaluation (as in the United States in early 1933); or, a banking crisis can cause devaluation, either through its adverse effects on aggregate demand or by affecting the supply of money (when a costly bank bail-out prompts monetization of government bail-out costs). Sovereign debt crises can also contribute to bank distress when banks hold large amounts of government debt (for example, in the banking crises in the United States in 1861, and in Argentina in 2001).

The consensus views regarding banking crises' origins (fundamental shocks versus confusion), the extent to which crises result from unwarranted runs on solvent banks, the social costs attending runs, and the appropriate policies to limit the costs of banking crises (government safety nets and prudential regulation) have changed dramatically, and more than once, over the course of the 19th and 20th centuries. Historical experience played a large role in changing perspectives toward crises, and the US experience had a disproportionate influence on thinking. Although panics were observed throughout world history (in Hellenistic Greece, and in Rome in AD 33), prior to the 1930s, in most of the world, banks were perceived as stable, large losses from failed banks were uncommon, banking panics were not seen as a great risk, and there was little perceived need for formal safety nets (for example, deposit insurance, or programmes to recapitalize banks). In many countries, ad hoc policies among banks, and sometimes including central banks, to coordinate bank responses to liquidity crises (as, for example, during the failure of Barings investment bank in London in 1890), seemed adequate for preventing systemic costs from bank instability.

Unusual Historical Instability of US Banks

The unusual experience of the United States was a contributor to changes in thinking that led to growing concerns about banks runs, and the need for aggressive safety net policies to prevent or mitigate runs. In retrospect, the extent to which US banking instability informed thinking and policy outside the United States seems best explained by the size and pervasive influence of the United States; in fact, the US crises were unique and reflected peculiar features of US law and banking structure.

The US panic of 1907 (the last of a series of similar US events, including 1857, 1873, 1884, 1890, 1893, and 1896) precipitated the creation of the Federal Reserve System in 1913 as a means of enhancing systemic liquidity, reducing the probability of systemic depositor runs, and mitigating the costs of such events. This innovation was specific to the United States (other countries either had established central banks long before, often with other purposes in mind, or had not established central banks), and reflected the unique US experience with panics – a phenomenon that the rest of the world had not experienced since 1866, the date of the last British banking panic (Bordo 1985).

For example, Canada did not suffer panics like those of the United States and did not establish a central bank until 1935. Canada's early decision to permit branch banking throughout the country ensured that banks were geographically diversified and thus resilient to large sectoral shocks (like those to agriculture in the 1920s and 1930s), able to compete through the establishment of branches in rural areas (because of the low overhead costs of establishing additional branches), and able to coordinate the banking system's response in moments of confusion to avoid depositor runs (the number of banks was small, and assets were highly concentrated in several nationwide institutions). Outside the United States, coordination among banks facilitated systemic stability by allowing banks to manage incipient panic episodes to prevent widespread bank runs. In Canada, the Bank of Montreal would occasionally coordinate actions by the large Canadian banks to stop crises before the public was even aware of a possible threat.

The United States, however, was unable to mimic this behaviour on a national or regional scale (Calomiris 2000; Calomiris and Schweikart 1991). US law prohibited nationwide branching, and most states prohibited or limited within-state branching. US banks, in contrast to banks elsewhere, were numerous (for example, numbering more than 29,000 in 1920), undiversified, insulated from competition, and unable to coordinate their response to panics (US banks established clearing houses, which facilitated local responses to panics beginning in the 1850s, as emphasized by Gorton 1985).

The structure of US banking explains why the United States uniquely had banking panics in which runs occurred despite the health of the vast majority of banks. The major US banking panics of the post-bellum era (listed above) all

occurred at business cycle peaks, and were preceded by spikes in the liabilities of failed businesses and declines in stock prices; indeed, whenever a sufficient combination of stock price decline and rising liabilities of failed businesses occurred, a panic *always* resulted (Calomiris and Gorton 1991). Owing to the US banking structure, panics were a predictable result of business cycle contractions that, in other countries, resulted in an orderly process of financial readjustment.

The United States, however, was not the only economy to experience occasional waves of bank failures before the First World War. Nor did it experience the highest bank failure rates, or bank failure losses. None of the US banking panics of the pre-First World War era saw nationwide banking distress (measured by the negative net worth of failed banks relative to annual GDP) greater than the 0.1 per cent loss of 1893. Losses were generally modest elsewhere, but Argentina in 1890 and Australia in 1893, where the most severe cases of banking distress occurred during this era, suffered losses of roughly ten per cent of GDP. Losses in Norway in 1900 were three per cent and in Italy in 1893 one per cent of GDP, but with the possible exception of Brazil (for which data do not exist to measure losses), there were no other cases in 1875–1913 in which banking loss exceeded one per cent of GDP.

Loss rates tended to be low because banks structured themselves to limit their risk of loss, by maintaining adequate equity-to-assets ratios, sufficiently low asset risk, and adequate asset liquidity. Market discipline (the fear that depositors would withdraw their funds) provided incentives for banks to behave prudently. The picture of small depositors lining up around the block to withdraw funds has received much attention, but perhaps the more important source of market discipline was the threat of an informed (often 'silent') run by large depositors (often other banks). Banks maintained relationships with each other through interbank deposits and the clearing of public deposits, notes and bankers' bills. Banks often belonged to clearing houses that set regulations and monitored members' behaviour. A bank that lost the trust of its fellow bankers could not long survive.

Changing Perceptions of Banking Instability

This perception of banks as stable, as disciplined by depositors and interbank arrangements to act prudently, and as unlikely to fail was common prior to the 1930s. The banking crises of the Great Depression changed this perception. US Bank failures resulted in losses to depositors in the 1930s in excess of three per cent of GDP. Bank runs, bank holidays (local and national government-decreed periods of bank closure to attempt to calm markets and depositors), and widespread bank closure suggested a chaotic and vulnerable system in need of reform. The Great Depression saw an unusual raft of banking regulations, especially in the United States, including restrictions on bank activities (the separation of commercial and investment banking, subsequently reversed in the 1980s and 1990s), targeted bank recapitalizations (the Reconstruction Finance Corporation), and limited government insurance of deposits.

Academic perspectives on the Depression fuelled the portrayal of banks as crisis-prone. The most important of these was the treatment of the 1930s banking crises by Milton Friedman and Anna Schwartz in their book, *A Monetary History of the United States* (1963). Friedman and Schwartz argued that many solvent banks were forced to close as the result of panics, and that fear spread from some bank failures to produce failures elsewhere. They saw the early failure of the Bank of United States in 1930 as a major cause of subsequent bank failures and monetary contraction. They lauded deposit insurance: 'federal deposit insurance, to 1960 at least, has succeeded in achieving what had been a major objective of banking reform for at least a century, namely, the prevention of banking panics'. Their views that banks were inherently unstable, that irrational depositor runs could ruin a banking system, and that deposit insurance was a success, were particularly influential coming from economists known for their scepticism of government interventions.

Since the publication of *A Monetary History of the United States*, however, other scholarship

(notably, the work of Wicker 1996 and Calomiris and Mason 1997, 2003a) has led to important qualifications of the Friedman–Schwartz view of bank distress during the 1930s, and particularly of the role of panic in producing distress. Detailed studies of particular regions and banks' experiences do not confirm the view that panics were a nationwide phenomenon during 1930 or early 1931, or an important contributor to nationwide distress until very late in the Depression (that is, early 1933). Regional bank distress was often localized and traceable to fundamental shocks to the values of bank loans. Not only does it appear that the failure of the Bank of United States had little effect on banks nationwide in 1930, one scholar has argued that there is evidence that the bank was, in fact, insolvent when it failed (Lucia 1985).

Other recent research on banking distress during the pre-Depression era has also de-emphasized inherent instability, and focused on the historical peculiarity of the US banking structure and panic experience, noted above. Furthermore, recent research on the destabilizing effects of bank safety nets has been informed by the experience of the US Savings and Loan industry debacle of the 1980s, the banking collapses in Japan and Scandinavia during the 1990s, and similar banking system debacles occurring in 140 developing countries in the last quarter of the 20th century, all of which experienced banking system losses in excess of one per cent of GDP, and more than 20 of which experienced losses in excess of ten per cent of GDP (data are from Caprio and Klingebiel 1996, updated in private correspondence with these authors). Empirical studies of these unprecedented losses concluded that deposit insurance and other policies that protect banks from market discipline, intended as a cure for instability, have become instead the single greatest source of banking instability.

The theory behind the problem of destabilizing protection has been well known for over a century, and was the basis for US President Franklin Roosevelt's opposition to deposit insurance in 1933 (an opposition shared by many). Deposit insurance was seen as undesirable special interest legislation designed to benefit small banks. Numerous attempts to introduce it failed to attract support in Congress (Calomiris and White 1994). Deposit insurance removes depositors' incentives to monitor and discipline banks, and frees bankers to take imprudent risks (especially when they have little or no remaining equity at stake, and see an advantage in 'resurrection risk taking'). The absence of discipline also promotes banker incompetence, which leads to unwitting risk taking.

Empirical research on late 20th-century banking collapses has produced a consensus that the greater the protection offered by a country's bank safety net, the greater the risk of a banking collapse (see, for example, Caprio and Klingebiel 1996, and the papers from a 2000 World Bank conference on bank instability listed in the bibliography). Empirical research on prudential bank regulation emphasizes the importance of subjecting some bank liabilities to the risk of loss to promote discipline and limit risk taking (Shadow Financial Regulatory Committee 2000; Mishkin 2001; Barth et al. 2006).

Studies of historical deposit insurance reinforce these conclusions (Calomiris 1990). The basis for the opposition to deposit insurance in the 1930s was the disastrous experimentation with insurance in several US states during the early 20th century, which resulted in banking collapses in all the states that adopted insurance. Government protection had played a similarly destabilizing role in Argentina in the 1880s (leading to the 1890 collapse) and in Italy (leading to its 1893 crisis). In retrospect, the successful period of US deposit insurance, from 1933 to the 1960s, to which Friedman and Schwartz referred, was an aberration, reflecting limited insurance during those years (insurance limits were subsequently increased), and the unusual macroeconomic stability of the era.

Models of banking crises followed trends in the empirical literature. The understanding of bank contracting structures, in light of potential crises, has been a consistent theme. Banks predominantly hold illiquid assets ('opaque,'

non-marketable loans), and finance those assets mainly with deposits withdrawable on demand. Banks are not subject to bankruptcy preference law, but rather, apply a first-come, first-served rule to failed bank depositors (depositors who are first in line keep the cash paid out to them). These attributes magnify incentives to run banks. An early theoretical contribution, by Diamond and Dybvig (1983), posited a banking system susceptible to the constant threat of runs, with multiple equilibria, where runs can occur irrespective of problems in bank portfolios or any fundamental demand for liquidity by depositors. They modelled deposit insurance as a means of avoiding the bad (bank run) equilibrium. Over time, other models of banks and depositor behaviour developed different implications, emphasizing banks' abilities to manage risk effectively, and the beneficial incentives of demand deposits in motivating the monitoring of banks in the presence of illiquid bank loans (Calomiris and Kahn 1991).

The literatures on banking crises also rediscovered an older line of thought emphasized by John Maynard Keynes (1931) and Irving Fisher (1933): market discipline implies links between increases in bank risk, depositor withdrawals and macroeconomic decline. As banks respond to losses and increased risk by curtailing the supply of credit, they can aggravate the cyclical downturn, magnifying declines in investment, production, and asset prices, whether or not bank failures occur (Bernanke 1983; Bernanke and Gertler 1990; Calomiris and Mason 2003b; Allen and Gale 2004; Von Peter 2004; Calomiris and Wilson 2004). New research explores general equilibrium linkages among bank credit supply, asset prices and economic activity, and adverse macroeconomic consequences of 'credit crunches' that result from banks' attempts to limit their risk of failure. This new generation of models provides a rational-expectations, 'shock-and-propagation' approach to understanding the contribution of financial crises to business cycles, offering an alternative to the endogenous-cycles, myopic-expectations view pioneered by Hyman Minsky (1975) and Charles Kindleberger (1978).

See Also

- ▶ Credit Rationing
- ▶ Currency Crises
- ▶ Deposit Insurance
- ▶ Great Depression
- ▶ Moral Hazard

Bibliography

Allen, F., and D. Gale. 2004. Financial fragility, liquidity, and asset prices. *Journal of the European Economic Association* 2: 1015–1048.

Barth, J.R., G. Caprio, and R. Levine. 2006. *Rethinking bank regulation: Till angels govern*. Cambridge: Cambridge University Press.

Bernanke, B.S. 1983. Nonmonetary effects of the financial crisis in the propagation of the great depression. *American Economic Review* 73: 257–276.

Bernanke, B.S., and M. Gertler. 1990. Financial fragility and economic performance. *Quarterly Journal of Economics* 105: 87–114.

Bordo, M. 1985. The impact and international transmission of financial crises: Some historical evidence, 1870–1933. *Revista di Storia Economica* 2(2d): 41–78.

Boyd, J., P. Gomis, S. Kwak, and B. Smith. 2000. *A user's guide to banking crises*. Conference paper. Washington, DC: World Bank.

Bruner, R.F., and S.D. Carr. 2007. *Money Panic: Lessons from the financial crisis of 1907*. New York: Wiley.

Calomiris, C.W. 1990. Is deposit insurance necessary? A historical perspective. *Journal of Economic History* 50: 283–295.

Calomiris, C.W. 2000. *U.S. bank deregulation in historical perspective.* Cambridge: Cambridge University Press.

Calomiris, C.W., and G. Gorton. 1991. The origins of banking panics: Models, facts, and bank regulation. In *Financial markets and financial crises*, ed. R.G. Hubbard. Chicago: University of Chicago Press.

Calomiris, C.W., and C.M. Kahn. 1991. The role of demandable debt in structuring optimal banking arrangements. *American Economic Review* 81: 497–513.

Calomiris, C.W., and J.R. Mason. 1997. Contagion and bank failures during the great depression: The June 1932 Chicago banking panic. *American Economic Review* 87: 863–883.

Calomiris, C.W., and J.R. Mason. 2003a. Fundamentals, panics and bank distress during the depression. *American Economic Review* 93: 1615–1647.

Calomiris, C.W., and J.R. Mason. 2003b. Consequences of bank distress during the great depression. *American Economic Review* 93: 937–947.

Calomiris, C.W., and L. Schweikart. 1991. The panic of 1857: Origins, transmission, and containment. *Journal of Economic History* 51: 807–834.
Calomiris, C.W., and E.N. White. 1994. The origins of federal deposit insurance. In *The regulated economy: A historical approach to political economy*, ed. C. Goldin and G. Libecap. Chicago: University of Chicago Press.
Calomiris, C.W., and B. Wilson. 2004. Bank capital and portfolio management: The 1930s 'capital crunch' and scramble to shed risk. *Journal of Business* 77: 421–455.
Caprio, G., and D. Klingebiel. 1996. *Bank insolvencies: Cross country experience*. Working paper No. 1620. Washington, DC: World Bank.
Cull, R., L. Senbet, and M. Sorge. 2000. *Deposit insurance and financial development*. Conference paper. Washington, DC: World Bank.
Demirguc-Kunt, A., and E. Detragiache. 2000. *Does deposit insurance increase banking system stability?* Conference paper. Washington, DC: World Bank.
Demirguc-Kunt, A., and H. Huizinga. 2000. *Market discipline and financial safety net design*. Conference paper. Washington, DC: World Bank.
Diamond, D., and P. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91: 401–419.
Fisher, I. 1933. The debt deflation theory of great depressions. *Econometrica* 1: 337–357.
Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.
Gorton, G. 1985. Clearing houses and the origin of central banking in the United States. *Journal of Economic History* 45: 277–283.
Honohan, P., and D. Klingebiel. 2000. *Controlling fiscal costs of banking crises*. Conference paper. Washington, DC: World Bank.
Keynes, J.M. 1931. The consequences to the banks of the collapse of money values. In *Essays in persuasion.* New York: W.W. Norton, 1963.
Kindleberger, C.P. 1978. *Manias, panics, and crashes: A history of financial crises*. New York: Basic Books.
Lucia, J.L. 1985. The failure of the bank of United States: A reappraisal. *Explorations in Economic History* 22: 402–416.
Minsky, H.P. 1975. *John Maynard Keynes*. New York: Columbia University Press.
Mishkin, F.S. 2001. *Prudential supervision: What works and what doesn't*. Chicago: University of Chicago Press.
Shadow Financial Regulatory Committee. 2000. *Reforming bank capital regulation*. Washington, DC: American Enterprise Institute.
Von Peter, G. 2004. *Asset prices and banking distress: A macroeconomic approach*. BIS working paper No. 167. Basel: Bank of International Settlements.
Wicker, E. 1996. *The banking panics of the great depression*. Cambridge: Cambridge University Press.

Banking Industry

Dario Focarelli and Alberto Franco Pozzolo

B

Keywords

Adverse selection; Assets and liabilities; Asymmetric information; Bank deregulation; Banking crises; Banking industry; Bankruptcy; Barriers to entry; Capital controls; Credit; Excessive risk taking; Glass–Stegall Act 1933 (USA); Information costs; Information monopolies; Liquidity; Mergers and acquisitions; Moral hazard; Relationship lending; Technological innovation

JEL Classifications

L66

The distinctive function of banks is the transformation of short-term deposits into longer-term, less liquid and riskier loans (Fama 1980, 1985; Diamond and Rajan 2001; Gorton and Winton 2003). By raising funds from depositors and providing credit, banks avoid the duplication of monitoring, which reduces the overall cost of transferring funds from capital suppliers to its users (Leland and Pyle 1977; Diamond 1984). At the same time, however, the greater liquidity of liabilities than of assets, which are typically longer-term and riskier, makes bank balance sheets vulnerable. Not only may banks fail if they are unable to obtain repayment of their loans, but depositors might even decide to withdraw their assets simply anticipating that others will do so. Such a 'bank run' can drive an otherwise sound bank to insolvency (Diamond and Dybvig 1983). The need to protect depositors and so guarantee a stable monetary transaction system explains why the banking industry is so heavily regulated. It is harder for a depositor to protect his interests than for an average investor, because judging the financial condition of a bank

is difficult and costly, even for specialists. For this reason, the typical instruments adopted by bank regulators include restrictions on the amount of risk that a bank can take, and compulsory deposit insurance schemes that prevent runs.

Regulatory intervention affects the shape of the banking industry and its degree of competition. Until the mid-1960s, governments deliberately limited competition in the interest of 'safety and soundness' by regulating deposit rates, entry, branching and mergers. The traditional view is of a trade-off between soundness and competition, with more intense competition reducing franchise values and increasing incentives to take on risky projects, since forgone future profits in the case of bankruptcy are lower (Keeley 1990). By increasing the equity at risk, capital controls reduce (although perhaps not entirely) excessive risk-taking (Hellman et al. 2000).

Recently, a more comprehensive view has been put forward, suggesting that regulation interacts dynamically with pervasive information asymmetries, and that the relationship between competition and stability is accordingly complex and multifaceted (Allen and Gale 2003). The cost of acquiring information in order to mitigate moral hazard and adverse selection is a strong endogenous barrier to the entry of new banks, allowing incumbents to gain monopoly rents (Broecker 1990), making competitive equilibria unsustainable (Dell'Ariccia 2001; Dell'Ariccia et al. 1999), and forcing new entrants to take a higher-risk clientele (Shaffer 1998).

The problems of information asymmetries can be attenuated if a bank deals repeatedly with the same customer, a practice known as 'relationship lending'. However, as Sharpe (1990) and Rajan (1992) show, this gives relationship banks a monopoly on information about their borrowers, further reducing competition, especially in the short run (Petersen and Rajan 1995). In this case, deregulation aimed at fostering inter-bank competition in transaction lending could have the effect of augmenting the scope for relationship banking, which permits banks to retain some monopoly power. As Boot and Thakor (2000) show, this is not the case if stronger competition comes from capital market financing, which drives some banks out of the market, reducing competition and consequently relationship lending.

Since the mid-1980s, the banking industry has been transformed by a series of events: deregulation of deposit accounts, which forced US banks to compete on interest rates; branching liberalization, which led to a sharp decline in the number of banks; the changes in capital requirements introduced with the Basel accords of 1988, which pushed banks towards newer and less regulated off-balance-sheet activities; the introduction of the euro, which created a unique wholesale banking market within Europe (Berger et al. 1995); and the substantial repeal of the Glass–Steagall Act of 1933, allowing banks to supply financial services previously offered only by other intermediaries, such as investment firms and insurance companies.

One of most important consequences of deregulation has been the unprecedented numbers of mergers and acquisitions during the 1990s, which sharply reduced the number of banks in many industrial countries and often heightened concern over possible anti-competitive effects. However, there is no clear evidence that the consolidations have harmed consumers or diminished competition, as would have been predicted from the observed negative correlation between the degree of concentration in local banking markets and the level of deposit rates (Berger and Hannan 1989). Rather, the available evidence indicates a positive effect stemming from the larger and more efficient banks taking over the smaller and less efficient (Berger et al. 1995; Focarelli et al. 2002). And while there may be some contraction of credit to smaller clients due to consolidation, this effect appears to be largely offset by increased lending by other banks (Berger et al. 1998). Indeed, there is evidence that in the medium term mergers increase the efficiency of the target bank, benefiting depositors (Focarelli and Panetta 2003).

The future of the banking industry is likely be determined by the interaction of three major forces: international competition, innovation in information technology and regulation. At

present, all three factors are heightening competition in banking. International competition, while still limited, tends to display the same pattern as domestic consolidation, with larger and more efficient banks in more developed countries taking over less efficient banks in financially less developed areas (Focarelli and Pozzolo 2005). Technological innovation is lessening the importance of close lending relationships, enlarging the size of local credit markets and further reducing the role of small banks (Petersen and Rajan 2002). Worldwide regulatory systems are moving to allow more competition and to assign a more important role to market evaluation (Basel Committee on Banking Supervision 2005).

See Also

- ▶ Agency Problems
- ▶ Banking Crises
- ▶ Financial Intermediation
- ▶ Market Structure
- ▶ Merger Analysis (United States)
- ▶ Microcredit
- ▶ Payment Systems

Bibliography

Allen, F., and D. Gale. 2003. Competition and financial stability. *Journal of Money, Credit, and Banking* 36: 433–480.

Basel Committee on Banking Supervision. 2005. *International convergence of capital measurement and capital standards: A revised framework*. Basel: BIS.

Berger, A., and T. Hannan. 1989. The price–concentration relationship in banking. *Review of Economics and Statistics* 71: 291–299.

Berger, A., A. Kashyap, and J. Scalise. 1995. The transformation of the US banking industry: What a long trip it's been. *Brookings Papers on Economic Activity* 1995(2): 55–201.

Berger, A., A. Saunders, J. Scalise, and G. Udell. 1998. The effects of bank mergers and acquisitions on small business lending. *Journal of Financial Economics* 50: 187–229.

Boot, A., and A. Thakor. 2000. Can relationship banking survive competition? *Journal of Finance* 55: 679–713.

Broecker, T. 1990. Credit-worthiness tests and interbank competition. *Econometrica* 58: 429–452.

Dell'Ariccia, G. 2001. Asymmetric information and the structure of the banking industry. *European Economic Review* 45: 1957–1980.

Dell'Ariccia, G., E. Friedman, and R. Marquez. 1999. Adverse selection as a barrier to entry in the banking industry. *RAND Journal of Economics* 30: 515–534.

Diamond, D. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51: 393–414.

Diamond, D., and P. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91: 401–419.

Diamond, D., and R. Rajan. 2001. Liquidity risk, liquidity creation and financial fragility: A theory of banking. *Journal of Political Economy* 109: 287–327.

Fama, E. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6: 39–57.

Fama, E. 1985. What's different about banks? *Journal of Monetary Economics* 15: 29–34.

Focarelli, D., and F. Panetta. 2003. Are mergers beneficial to consumers? Evidence from the market for bank deposits. *American Economic Review* 93: 1152–1172.

Focarelli, D., and A. Pozzolo. 2005. Where do banks expand abroad? An empirical analysis. *Journal of Business* 78: 2435–2464.

Focarelli, D., F. Panetta, and C. Salleo. 2002. Why do banks merge? *Journal of Money, Credit, and Banking* 34: 784–803.

Gorton, G., and A. Winton. 2003. Financial intermediation. In *Handbook of the economics of finance*, ed. G. Constantinides, M. Harris, and R. Stulz, vol. 1. Amsterdam: North-Holland.

Hellman, T., K. Murdock, and J. Stiglitz. 2000. Liberalization, moral hazard in banking and prudential regulation: Are capital requirements enough? *American Economic Review* 90: 147–165.

Keeley, M. 1990. Deposit insurance, risk, and market power in banking. *American Economic Review* 80: 1183–1200.

Leland, H., and D. Pyle. 1977. Informational asymmetries, financial structure and financial intermediation. *Journal of Finance* 32: 371–387.

Petersen, M., and R. Rajan. 1995. The effect of credit market competition on lending relationships. *Quarterly Journal of Economics* 110: 407–443.

Petersen, M., and R. Rajan. 2002. Does distance still matter? The information revolution in small business lending. *Journal of Finance* 57: 2533–2570.

Rajan, R. 1992. Insiders and outsiders: The choice between relationship and arms length debt. *Journal of Finance* 47: 1367–1400.

Shaffer, S. 1998. The winner's curse in banking. *Journal of Financial Intermediation* 7: 359–392.

Sharpe, S. 1990. Asymmetric information, bank lending and implicit contracts: A stylized model of customer relationships. *Journal of Finance* 45: 1069–1087.

Banking School, Currency School, Free Banking School

Anna J. Schwartz

Abstract

The doctrines of the three nineteenth century schools differed. The Currency School believed that note issues should vary one-to-one with the Bank of England's gold reserves. The Banking School believed that real bills, needs of trade and the law of reflux should govern bank operations. The Free Banking School believed that competitive private banks would not overissue, whereas a monopoly issuer did so. Other issues were debated. Was a central bank needed? Should a central bank be subject to rules or allowed discretion? How should money be defined? No one point of view carried the day and several of the issues that divided the schools are still debated today.

Keywords

Balance of payments; Bank Charter Act 1833 (UK); Bank Charter Act 1844 (UK); Bank of England; Bank of Ireland; Bank of Scotland; Banking School; Bullion reserve; Central banking; Convertibility; Country banks; Credit; Currency principle; Currency School; Free Banking School; Fullarton, J.; Gilbart, J.; Gold standard; Inflation; Joint stock banking; Law of reflux; Longfield, M.; McCulloch, J. R.; Mill, J. S.; Money supply; Money, definition of; Monopoly of note issue; Needs of trade doctrine; Norman, G.; Note issue; Overissue; Overstone, Lord; Parnell, H.; Private banks; Real bill doctrine; Reserve-deposit ratio; Rules versus discretion; Schwartz, A. J.; Scrope, G. P.; Stocks and flows; Tooke, T.; Torrens, R.; Trade cycles; Wilson, J.

JEL Classifications

N1

Historians of economic thought conventionally represent British monetary debates from the 1820s on as centred on the question of whether policy should be governed by rules (espoused by adherents of the Currency School), or whether authorities should be allowed discretion (espoused by adherents of the Banking School). In fact many other questions were in dispute, including those raised by neglected or misidentified participants in the debates – adherents of the Free Banking School.

Among the questions in dispute were the following: (1) Should the banking system follow the Currency School's principle that note issues should vary one-to- one with the Bank of England's gold holdings? (2) Were the doctrines of the Banking School – real bills, needs of trade and the law of reflux – valid? (3) Was a monopoly of note issue desirable or, as the Free Banking School contended, destabilizing? (4) Was overissue a problem and, if so, who was responsible? (5) How should money be defined? (6) Why do trade cycles occur? (7) Should there be a central bank? No, was the Free Banking School answer to the final question; yes, was the answer of the other two schools, with disparate views, as indicated, on the question of rules vs. authorities. What was not in dispute was the viability of the gold standard system with gold convertibility of Bank of England notes.

On what grounds did the schools oppose each other? Each of the first three questions identifies the central doctrines that the adherents of one of the schools shared; on the remaining questions, individual views within each school varied. Before establishing the positions of each school in the monetary debates, we introduce the institutional background and the principal participants.

Institutional Background

The Bank of England, incorporated in 1694 as a private institution with special privileges, stood at the head of the British banking system at the time of the debates. Until 1826 the Bank's charter was interpreted to mean the prohibition of other joint stock banks in England. As a result banking establishments were either one-man firms or

partnerships with not more than six members. Two types of banks predominated in England: the wealthy London private banks which had voluntarily surrendered their note-issuing privilege, and the country banks which depended almost exclusively on the business of note issues. Numerous failures among the country banks demonstrated that the effect of the Bank's charter was to foster the formation of banking units of uneconomical size.

Banking in Ireland was patterned on English lines. The Bank of Ireland, chartered in 1783 with the exclusive privilege of joint stock banking in Ireland, surrendered its monopoly in 1821 in places farther than fifty miles from Dublin. Joint-stock banking in the whole of Ireland was legalized in 1845.

The Bank of Scotland was founded in 1695 with privileges similar to those of the Bank of England, except that it was formed to promote trade, not to support the credit of the government. It lost its monopoly in 1716, and no further monopolistic banking legislation was enacted in Scotland. With free entry possible, many local private and joint stock banks, most of the latter well capitalized, were established, and a nationwide system of branch banking developed. Unlike the English system, overissue was not a problem in the Scottish system. The banks accepted each other's notes and evolved a system of note exchange. Shareholders of Scottish joint stock banks (except for three chartered banks) assumed unlimited liability. At the time of the debates banking in Scotland was at a far more advanced stage than in England.

Principals in the Debates

The leading spokesmen for the Currency School side in the debates were McCulloch, Loyd (later Lord Overstone), Longfield, George Warde Norman, and Torrens. Norman, a director of the Bank of England for most of the years 1821–1872, and of the Sun Insurance Company, 1830–1864, was active in the timber trade with Norway. The principal Banking School representatives were Tooke, Fullarton, and John Stuart Mill, while James Wilson held views that straddled Banking and Free Banking School doctrines. The most prominent members of the Free Banking School were Parnell (later Baron Congleton), James William Gilbart, and Poulett Scrope. Gilbart, a banker, was general manager of the London and Westminster Bank, the first of the joint stock banks authorized by the Bank Charter Act of 1833.

Currency School Principle

The objective of the Currency School was to achieve a price level that would be the same whether the money supply were fully metallic or a mixed currency including both paper notes and metallic currency. According to Loyd, gold inflows or outflows under a fully metallic currency had the immediate effect of increasing or decreasing the currency in circulation, whereas a mixed currency could operate properly only if inflows or outflows of gold were exactly matched by an increase or decrease of the paper component. He and others of the Currency School regarded a rise in the price level and a fall in the bullion reserve under a mixed currency as symptoms of excessive note issues. They advocated statutory regulation to ensure that paper money was neither excessive nor deficient because otherwise fluctuations in the currency would exacerbate cyclical tendencies in the economy. They saw no need, however, to regulate banking activities other than note issue.

The Banking School challenged these propositions. Fullarton denied that overissue was possible in the absence of demand, that variations in the note issue could cause changes in the domestic price level, or that such changes could cause a fall in the bullion reserve ([1844] 1969, pp. 57, 128–129). Under a fully metallic as well as under a mixed currency bank, deposits, bills of exchange, and all forms of credit might influence prices. Moreover, inflows and outflows of gold under a fully metallic currency might change bullion reserves but not prices. If convertibility were maintained, overissue was not feasible and no statutory control of note issues was required. An adverse balance of payments was a temporary phenomenon that was self-correcting when, for

example, a good harvest followed a bad one. According to the Free Banking School, the possibility of overissue and inflation applied only to Bank of England notes but could not occur in a competitive banking system.

Banking School Principle

The Banking School adopted three principles that for them reflected the way banks actually operated as opposed to the Currency School principle which they dismissed as an artificial construct of certain writers (White 1984, pp. 119–128).

The first Banking School principle was the doctrine that liabilities of deposits and notes would never be excessive if banks restricted their earning assets to real bills. One charge levelled by modern economists against the doctrine is that it leaves the quantity of money and the price level indeterminate, since it links the money supply to the nominal magnitude of bills offered for discount. Some members of the school may be exculpated from this charge if they regarded England as a small open economy, its domestic money stock a dependent variable determined by external influences. However, because it ignored the role of the discount rate in determining the volume of bills generated in trade, the doctrine was vulnerable. In addition, the Banking School confused the flow demand for loanable funds, represented by the volume of bills, with the stock demand for circulating notes, although the two magnitudes are non-commensurable.

Free Banking School members who also adopted the real bills doctrine erroneously attributed overissue by the Bank of England to its purchase of assets other than real bills, when overissue was possible with a portfolio limited to real bills, acquired at an interest rate that led to a stock of circulating medium inconsistent with the prevailing price level (Gilbart 1841, pp. 103–105, 119–120). The Currency School regarded the real bills doctrine as misguided since it could promote a cumulative rise in the note issue and hence in prices.

A second Banking School principle was the 'needs of trade' doctrine, to the effect that the note circulation should be demand-determined – curtailed when business declined and expanded when business prospered, whether for seasonal or cyclical reasons. An implicit assumption of the doctrine was that banks could either vary their reserve ratios to accommodate lower or higher note liabilities, or else offset changes in note liabilities by opposite changes in deposit liabilities. For non-seasonal increases in demand for notes, the doctrine implied that expanding banks could obtain increased reserves from an interregional surplus of the trade balance. The Currency School regarded an increase in the needs of trade demand to hold notes accompanying increases in output and prices as unsound because it would ultimately produce an external drain. The Free Banking School countered that such an objection by the Currency School was paradoxical since the virtue of a metallic currency according to the latter was that it accommodated the commercial wants of the country, and therefore for a mixed currency to respond to the needs of trade could not be a vice. The modern objection to the needs of trade doctrine as procyclical is an echo of the Currency School view.

The third Banking School principle was the law of the reflux according to which overissue was possible only for limited periods because notes would immediately return to the issuer for repayment of loans. This was a modification of the real bills doctrine that Tooke and Fullarton advanced, since adherence to the doctrine supposedly made overissue impossible. They made no distinction between the speed of the reflux for the Bank of England and for competitive banks of issue – a distinction at the heart of the Free Banking position. For the latter, reflux of excess notes was speedy only if the notes were deposited in rival banks. These would then return the notes to the issuing banks and accordingly bring an end to relative overissue by individual banks. The Bank of England, on the contrary, could overissue for long periods because it had no rivals. Fullarton, however, made the unwarranted assumption that notes would be returned to the Bank to repay previous loans at a faster rate than the Bank was discounting new loans, hence correcting the overissue. Moreover, he believed that if the Bank overissued by open market

purchases, the decline in interest rates would quickly activate capital outflows, reducing the Bank's bullion and forcing it to retreat. Tooke was sounder in arguing for the law of reflux on the ground that excess issues would not be held if they did not match the preferences of holders for notes rather than deposits.

The Banking School had no legislative programme for reform of the monetary system. Good bank management, in the view of the school, could not be legislated.

Free Banking School Principle

As the name suggests, the principle the Free Banking School advocated was free trade in the issue of currency convertible into specie. Members of the school favoured a system like the Scottish banking system, where banks competed in all banking services, including the issue of notes, and no central bank held a monopoly of note issue. They argued that in such a system banks did not issue without limit but indeed provided a stable quantity of money, Although the costs of printing and issuing were minimal, to keep notes in circulation required restraint in their issue. The profit-maximizing course for competitive banks was to maintain public confidence in their issues by maintaining convertibility into specie on demand, which required limiting their quantity.

Loyd's response to the argument for free trade in currency was that unlike ordinary trades, what was sought was not the greatest quantity at the cheapest price but a regulated quantity of currency. The Free Banking School denied that free banking would debase the currency, and contended that the separation of banking from note issue, the Banking School proposal, was impractical. Scrope (1833a, pp. 32–33) asked why the Currency School objected to unregulated issue of notes but not to that of deposits, questioning Loyd's assumption that an issuing bank's function was to produce money, when in fact its function was to substitute its bank notes for less well-known private bills of exchange that were the bank's assets.

Scrope and other Free Banking adherents (Parnell 1827, p. 143) neglected the distinction between a banknote immediately convertible into gold and a commercial bill whose present value varied with time to maturity and the discount rate. Contrary to Loyd, they reasoned that free trade and competition were applicable to currency creation because the business of banks was to produce the scarce good of reputation.

Loyd's second disagreement with the argument for free trade in banking was that miscalculations by the issuers were borne not by them but by the public. Moreover, individuals had no choice but to accept notes they received in ordinary transactions, and trade in general suffered as a result of overissue. The Free Banking School answer to this externalities argument turned on the ability of holders to refuse notes of issuers without reputation. Protection against loss could also be provided if joint stock banks were allowed to operate in place of country banks limited to six or fewer partners. In addition, if banks were required to deposit security of government bonds or other assets, noteholders would be further protected (Scrope 1832, p. 455; 1833b, p. 424; Parnell 1827, pp. 140–144). Free Banking School members who argued in this vein failed to recognize that they were thereby acknowledging a role for government intervention in currency matters.

In the 1820s the Free Banking School championed joint stock banking both in the country bank industry and in direct competition in note issue with the Bank of England in London. Although the six-partner rule for banks of issue at least 65 miles from London was repealed in 1826 after a spate of bank failures, the Bank retained its monopoly of note circulation in the London area. In addition, the Bank was permitted to establish branches anywhere in England. The Parliamentary inquiry in 1832 on renewal of the Bank's character was directed to the question of prolonging the monopoly. The Act of 1833 eased entry for joint stock banks within the 65-mile limit but denied them the right of issue and made the Bank's notes legal tender for redemption of country bank notes, in effect securing the Bank's monopoly. The doom of the Free Banking cause was finally pronounced by the Bank Charter Act

of 1844. It restricted note issues of existing private and joint stock banks in England and Wales to their average circulation during a period in 1843. Note issue by banks established after the Act was prohibited.

Was Overissue a Problem?

Participants in the debates understood overissue to mean a stock of notes, whether introduced by a single issuer or banks in aggregate, in excess of the quantity holders voluntarily chose to keep as assets, given the level of prices determined by the world gold standard. Was overissue of a convertible currency possible? According to the Free Banking School, interbank note clearing by competitive banks operated to eliminate excess issued by a single bank. The check to excess issues by the banking system as a whole was an external drain through the price-specie flow mechanism. In this respect the school acknowledged that the result of overissue by a competitive banking system as a whole was the same as for a monopoly issuer. However, they held that overissue was a phenomenon that the monopoly of the Bank of England encouraged but a competitive system would discourage.

The Currency School, on the other hand, regarded both the Bank of England and the Scottish and country banks as equally prone to overissue and did not grant that a check to overissue by a single bank or banks in the aggregate was possible through the interbank note clearing mechanism. For them, regulation of a monopoly issuer promised a stable money supply that was not attainable with a plural banking system.

The Free Banking School's explanation of the Bank of England's ability to overissue rested on the absence of rivals for the Bank's London circulation, so no interbank note clearing took place; the absence of competition in London from interest-bearing demand deposits; and the fact that London private banks held the Bank's notes as reserves. Hence the demand for its notes was elastic. The Free Banking and Currency Schools agreed that there was a substantial delay before an external drain checked overissue, so the Bank's actions inescapably inflicted damage on the economy. Scrope (1830, pp. 57–60), who attributed the Bank's willingness to overexpand its note issues to its monopoly position, advocated abrogating that legal status.

The Banking School dismissed the question of overissue as irrelevant, for noteholders could easily exchange unwanted notes by depositing them. What they failed to examine was the possibility that a broader monetary aggregate could be in excess supply resulting in an external drain.

How Should Money Be Defined?

Currency School members favoured defining money as the sum of metallic money, government paper money, and bank notes (Norman 1833, pp. 23, 50; McCulloch 1850, pp. 146–147). The Free Banking School, like the Currency School, focused on bank notes as the common medium of exchange, ignoring demand deposits that were not usually subject to transfer by check outside London. The Banking School definition of money is sometimes represented as broader than that of the other schools, but in fact was narrower – money was restricted to metallic and government paper money. Bank notes and deposits were excluded, since they were regarded as means of raising the velocity of bank vault cash but not as adding to the quantity of money (Tooke [1848] 1928, pp. 171–183; Fullarton [1844] 1969, pp. 29–36; Mill [1848] 1909, p. 523). In the short run, the school held that all forms of credit might influence prices, but only money as defined could do so in the long run, because the domestic price level could deviate only temporarily from the world level of prices determined by the gold standard.

Why Do Trade Cycles Occur?

The positions of the three schools on the impulses initiating trade cycles were not dogma for their members. In general the Currency and Banking

Schools held that nonmonetary causes produced trade cycles, whereas the Free Banking School pointed to monetary causes, but individual members did not invariably hew to these analytical lines. McCulloch (1837, p. 63), Loyd (1857, p. 317), and Longfield (1840, pp. 222–223) essentially attributed cycles to waves of optimism and pessimism to which the banks then responded by expanding and contracting their issues. Banks accordingly never initiated the sequence of expansion and contraction. Hence the Currency School principle of regulating the currency to stabilize prices and business did not imply that cycles would thereby be eliminated. Cycles would, however, no longer be amplified by monetary expansion and contraction, if country banks were denied the right to issue and the Bank of England's circulation were governed by the 'currency principle'. Torrens (1840, pp. 31, 42–43), unlike other Currency School members, attributed trade cycles to actions of the Bank of England. That was also the position of the Free Banking School, although in an early work Parnell (1827, pp. 48–51) of that school held that cycles were caused by nonmonetary factors. For the Banking School, however, nonmonetary factors accounted for both the origin and spread of trade cycles. Tooke (1840, pp. 245, 277), for example, believed that overoptimism would prompt an expansion of trade credit for which the banks were in no way responsible. Collapse of optimism would then lead to shrinkage of trade credit. For Fullarton ([1844] 1969, p. 101) nonmonetary causes produced price fluctuations to which changes in note circulation were a passive response. Proponents of the nonmonetary theory of the onset of trade cycles provided no explanation of the waves of optimism and pessimism themselves. For the Free Banking School the waves were precipitated by the Bank of England's expansion and ultimate contraction of its liabilities. Initially, the Bank's actions depressed interest rates and ultimately forced them up, as loanable funds increased in supply and then decreased. The Bank's monopoly position enabled it to create such monetary disturbances, whereas competitive country banks had no such power.

Should There Be a Central Bank?

The Currency and Banking Schools were in agreement that a central bank with the sole right of issue was essential for the health of the economy. McCulloch (1831, p. 49) regarded a system of competitive note issuing institutions as one of inherent instability. Tooke (1840, pp. 202–207) favoured a monopoly issuer as promoting less risk of overissue and greater safety because it would hold sufficient reserves. The two schools differed on the need for a rule to regulate note issues, the Currency School pledged to a rulebound authority, the Banking School to an unbound authority. The Free Banking School disapproved of both a rule and a central bank authority, instead favouring a competitive note-issuing system that it held to be self-regulating. For that school proof that centralized power was inferior to a competitive system was revealed by cyclical fluctuations that had been caused by errors of the Bank of England.

A Continuing Debate

The Bank Charter Act of 1844 ended the right of note issue for new banks in England and Wales. Scottish banks, however, were treated differently from Irish banks by the Act of 1845 and from English provincial banks by the Act of 1844. Like the latter, authorized circulation for the Scottish banks was determined by the average of a base period, but they could exceed the authorized circulation provided they held 100 per cent specie reserves against the excess – a provision also imposed on the Bank of England.

The Free Banking School thus lost its case for an end of the note issue monopoly of the Bank of England. The death of Parnell in 1842, a leading Parliamentary spokesman, had hurt the cause. Others of the school were mainly country and joint stock bankers. The Acts conferred benefits on them by restricting entry into the note-issuing industry and by freezing market shares (White 1984, pp. 78–79). Their voices were not raised in opposition. Only Wilson was critical of the

privileges the Bank of England was accorded ([1847] 1859, pp. 34–66).

The Banking School objected not only to the Act but claimed vindication for its point of view by the necessity to suspend it in 1847, 1857 and 1866. The Currency School responded that the suspensions were of no great significance (Loyd 1848, pp. 393–394). The recommendations of the Currency School prevailed to set a maximum for country bank note issues and the eventual transfer of their circulation to the Bank of England.

The monetary debates that were initiated in the 1820s were not conclusive. No point of view carried the day. Long after the original participants had passed from the scene, the doctrines of the schools found supporters. Even the Free Banking School position in opposition to monopoly issue of hand-to-hand currency that seemed to be buried has recently been revived by new adherents (White 1984, pp. 137–150). The debate on all the questions in dispute in the 19th century continues to be live.

See Also

- ▶ Boyd, Walter (1754–1837)
- ▶ Bullionist Controversies (Empirical Evidence)
- ▶ Fullarton, John (1780–1849)
- ▶ Money, Classical Theory of
- ▶ Overstone, Lord [Samuel Jones Loyd] (1796–1883)
- ▶ Real Bills Doctrine
- ▶ Tooke, Thomas (1774–1858)

Bibliography

Fullarton, J. 1844. *On the regulation of currencies*. London: John Murray. Reprinted, New York: Augustus M. Kelley, 1969.

Gilbart, J.W. 1841. Testimony before the Select Committee of the House of Commons on Banks of Issue. British Sessional Papers, vol. 5 (410).

Gregory, T.E. 1928. *Introduction to Tooke and Newmarch's a history of prices*. London: P.S. King.

Longfield, S.M. 1840. Banking and currency. *Dublin University Magazine.*

Loyd, S.J. 1848. Testimony before the Secret Committee of the House of Commons on Commercial Distress. British Sessional Papers, 1847–8, vol. 8, part 1 (584).

Loyd, S.J. 1857. *Tracts and other publications on metallic and paper money*. London.

McCulloch, J.R. 1831. *Historical sketch of the Bank of England*. London: Longman.

McCulloch, J.R. 1837. The Bank of England and the country banks. *Edinburgh Review.*

McCulloch, J.R. 1850. *Essays on interest, exchange, coins, paper money, and banks*. London.

Mill, J.S. 1848. *Principles of political economy,* ed. W.J. Ashley. Reprinted London: Longmans & Co., 1909.

Norman, G.W. 1833. *Remarks upon some prevalent errors, with respect to currency and banking*. London: Hunter.

Parnell, H.B. 1827. *Observations on paper money, banking and overtrading*. London: James Ridgway.

Scrope, G.P. 1830. *On credit-currency, and its superiority to coin, in support of a petition for the establishment of a cheap, safe, and sufficient circulating medium*. London: John Murray.

Scrope, G.P. 1832. The rights of industry and the banking system. *Quarterly Review,* 407–455.

Scrope, G.P. 1833a. *An examination of the bank charter question*. London: John Murray.

Scrope, G.P. 1833b. *Principles of political economy.* London: Longman.

Tooke, T. 1840. *A history of prices and of the state of the circulation in 1838 and 1839*. London: Longman. Reprinted, London: P.S. King, 1928.

Tooke, T. 1848. *History of prices and of the state of the circulation, from 1839 to 1847 inclusive*. London: Longmans. Reprinted, London: P.S. King, 1928.

Torrens, R. 1840. *A letter to Thomas Tooke, Esq. in reply to his objections against the separation of the business of the bank into a Department of Issue and a Department of Discount: With a plan of bank reform*. London: Longman.

White, L.H. 1984. *Free banking in Britain: Theory, experience, and debate, 1800–1845*. Cambridge: Cambridge University Press.

Wilson, J. 1847. *Capital, currency, and banking; being a collection of a series of articles published in the Economist in 1845 ... and in 1847*. 2nd ed. London: The office of the Economist. London: D.M. Aird, 1859.

Bankruptcy Law, Economics of Corporate and Personal

Michelle J. White

Abstract

Bankruptcy is the legal process whereby financially distressed firms, individuals, and occasionally governments resolve their debts. The

bankruptcy process for firms plays a central role in economics, because competition tends to drive inefficient firms out of business, thereby raising the average efficiency level of those remaining. Bankruptcy also has an important economic function for individual debtors, since it provides them with partial consumption insurance and supplements the government-provided safety net. This article discusses the economic objectives of bankruptcy and surveys theoretical and empirical research on corporate and personal bankruptcy.

Keywords

Absolute Priority Rule (APR); Auctions; Bankruptcy; Bankruptcy contracting; Bankruptcy law, economics of corporate and personal; Bankruptcy, economics of; Consumption insurance; Equity finance; Fresh start; Limited liability; Liquidation; Options; Reorganization; Risk and return; Strategic default

JEL Classifications

K3; K35; K2; G3

Bankruptcy is the legal process whereby financially distressed firms, individuals, and occasionally governments resolve their debts. The bankruptcy process for firms plays a central role in economics, because competition tends to drive inefficient firms out of business, thereby raising the average efficiency level of those remaining. Consumers benefit because the remaining firms produce goods and services at lower costs and sell them at lower prices. The legal mechanism through which most firms exit the market is bankruptcy. Bankruptcy also has an important economic function for individual debtors, since it provides them with partial consumption insurance and supplements the government-provided safety net. Local governments occasionally also use bankruptcy to resolve their debts, and there has been discussion of establishing a bankruptcy procedure for financially distressed countries (see White 2002).

Bankruptcy Law

For both corporate and individual debtors, bankruptcy law provides a collective framework for simultaneously resolving all debts when debtors' assets are less valuable than their liabilities. This includes both rules for determining which of the debtor's assets must be used to repay debt and rules for dividing the assets among creditors. Thus bankruptcy is concerned with both the size of the pie – the total amount paid to creditors – and how the pie is divided.

For financially distressed corporations, both the size and the division of the pie depend on whether the corporation liquidates or reorganizes in bankruptcy, and bankruptcy law also includes rules for deciding whether reorganization or liquidation will occur. When corporations liquidate under Chap. 7 of US bankruptcy law, the pie includes all of the firm's assets but none of its owners' other assets. This reflects the doctrine of limited liability, which exempts owners of equity in corporations from personal liability for the corporation's debts beyond loss of the value of their shares. The corporation's assets are liquidated and the proceeds are used to repay creditors according to the absolute priority rule (APR). The APR carries into bankruptcy the non-bankruptcy rule that debt must be repaid in full before equity receives anything. The APR also determines how the pie is divided among creditors. Classes of creditors are ranked and each class receives full payment of its claims until funds are exhausted.

When corporations reorganize under Chap. 11 of US bankruptcy law, the reorganized corporation retains most or all of its assets and continues to operate – generally under the control of its pre-bankruptcy managers. Bankruptcy law again provides a procedure for determining both the size and the division of the pie in reorganization, but the procedure involves a negotiation process rather than a formula.

Funds to repay creditors come from the firm's future earnings rather than from liquidating its assets. The rule for the division of the pie in reorganization is also different. Instead of creditors receiving either full payment or nothing, most classes of creditors receive partial payment regardless of their

rank, and pre-bankruptcy equity receives some of the reorganized firm's new shares. This priority rule is referred to as 'deviations from the APR' since equity receives a positive payoff even though creditors are repaid less than 100 per cent. Creditors and equity negotiate a reorganization plan that specifies what each group will receive, and the plan must be adopted by a super-majority vote of each class of creditors and equity.

For individuals in financial distress, bankruptcy law also includes both rules for determining which of the individual's assets must be used to repay debt (the size of the pie) and rules for dividing the assets among creditors (the division of the pie). In determining the size of the pie, personal bankruptcy law plays a role similar to that of limited liability for corporate equity-holders, since it limits the amount of assets that individual debtors must use to repay. It does this by specifying exemptions, which are maximum amounts of both financial wealth and post-bankruptcy earnings that individual debtors are allowed to keep. Only amounts in excess of the exemption levels must be used to repay. An important feature of US bankruptcy law is the 100 per cent exemption for post-bankruptcy earnings, known as the 'fresh start', which greatly limits individual debtors' obligation to repay. (Note that in 2005 Congress adopted limits on the availability of the fresh start.) In personal bankruptcy, the rule for dividing repayment among creditors is also the APR.

An important difference between personal and corporate bankruptcy law is that, while corporations may either liquidate or reorganize in bankruptcy, individuals can only reorganize (even though the most commonly used personal bankruptcy procedure in the United States is called liquidation). This is because part of individual debtors' wealth is their human capital, and the only way to liquidate human capital is to sell debtors into slavery – as the Romans did. Since slavery is no longer used as a penalty for bankruptcy, all personal bankruptcy procedures are forms of reorganization in which individual debtors keep their human capital and the right to decide whether to use it.

Economic Objectives

The economic objectives are similar in corporate and personal bankruptcy. One important objective of bankruptcy is to require sufficient repayment that lenders will be willing to lend – not necessarily to the bankrupt debtor but to other borrowers. Reduced access to credit makes debtors worse off because businesses need to borrow in order to grow and individuals benefit from borrowing to smooth consumption. On the other hand, repaying more to creditors harms debtors by making it more difficult for financially distressed firms to survive and by reducing financially distressed individuals' incentive to work. Both the optimal size and the division of the pie in bankruptcy are affected by this trade-off. A second important objective of both types of bankruptcy is to prevent creditors from harming debtors by racing to be first to collect. When creditors think that a debtor is in financial distress, they have an incentive to collect their debts quickly, since the debtor will be unable to repay all creditors in full. But aggressive collection efforts by creditors may force debtor firms to shut down even when the best use of their assets is to continue operating, and may cause individual debtors to lose their jobs (if creditors repossess their cars or garnish their wages). A third objective of personal bankruptcy law that has no counterpart in corporate bankruptcy is to provide individual debtors with partial consumption insurance. If consumption falls substantially, long-term harm may occur, including debtors' children leaving school prematurely in order to work or debtors' medical conditions going untreated and becoming disabilities. Discharging debt in bankruptcy when debtors' consumption would otherwise fall reduces these costs. An additional objective that applies only to corporate bankruptcy is to reduce filtering failure. Financially distressed firms may be economically either efficient or inefficient, depending on whether the best use of their assets is the current use or some alternative use. Filtering failure in bankruptcy occurs when efficient but financially distressed firms shut down and when inefficient financially distressed firms reorganize and continue

operating. The cost of filtering failure is either that the firm's assets remain tied up in an inefficient use or that they move to an alternative use when the current one is the most efficient. Many researchers have argued that reorganization in Chap. 11 tends to save economically inefficient firms that should shut down.

Research on corporate and personal bankruptcy is discussed separately below. Small-business bankruptcy is included with personal bankruptcy, because small businesses are often unincorporated and therefore their debts are legal liabilities of the business owner. When these businesses fail, their owners can file for bankruptcy and both their business and personal debts will be discharged. (Note that most of the research on bankruptcy is focused on US law and US data. For a longer survey of research on corporate and personal bankruptcy that includes many references, see White 2006.)

Corporate Bankruptcy

Theory

A central theoretical question in corporate bankruptcy is how priority rules affect the efficiency of decisions made by managers (who are assumed to represent the interests of equity), particularly whether the firm invests in safe or risky projects and whether and when it files for bankruptcy. Inefficient investment decisions lower the firm's return, and inefficient bankruptcy decisions result in filtering failure. Both reduce creditors' returns and cause them to raise interest rates or to reduce the amount they are willing to lending.

Bebchuk (2002) compares the efficiency of corporate investment decisions when the priority rule in bankruptcy is the APR with those when deviations from the APR occur, where use of the APR represents liquidation in bankruptcy and deviations from the APR represent reorganization in bankruptcy. A well-known result in finance is that equity prefers risky to safe investment projects, because equity gains disproportionately when risky projects succeed and bears only limited losses when risky projects fail. If the priority rule in bankruptcy is changed from the APR to deviations from the APR, then equity's preference for risky projects becomes even stronger. This is because equity now receives a positive return rather than nothing when risky projects fail, and the same high return when risky projects succeed. This change makes risky projects even more attractive relative to safe ones, since the latter rarely fail and so their return is unaffected by the change in the priority rule. Thus, when the bankruptcy regime is reorganization rather than liquidation, investment decisions become less efficient because equity over-invests in risky projects.

But Bebchuk argues that the results are reversed when firms are already in financial distress. Here, deviations from the APR reduce rather than increase equity's bias towards choosing risky investment projects. This is because, when the project is likely to fail and the firm to file for bankruptcy, equity's main return comes from the share that it receives of the firm's value in bankruptcy – the deviations from the APR. And since safe projects have higher downside returns, they generate more for equity. Thus the overall result is that neither priority rule in bankruptcy always leads to efficient investment incentives. Similar models have shown that none of the standard priority rules always leads to efficient bankruptcy decisions.

Bankruptcy law also affects other economically important decisions, including whether managers default strategically, whether they reveal important information about the firm's condition to creditors, and how much effort they expend. Strategic default occurs when firms default on their debt even though they are financially solvent. In the financial contracting literature, there is a trade-off between strategic default and filtering failure (see Bolton and Scharfstein 1996). Suppose a firm borrows D in period 0 to finance an investment project. The firm will either succeed or fail. If it succeeds, it earns $R_1 > D$ in period 1 and an additional $R_2 > L$ in period 2. If it fails, then its period 1 earnings are zero, but it still earns R_2 in period 2. Regardless of whether

the firm succeeds or fails, the liquidation value of its assets is L in period 1 and 0 in period 2. The firm's earnings are assumed to be observable but unverifiable. The loan contract calls for the firm to repay D in period 1 and it gives lenders the right to liquidate the firm in period 1 and collect L if default occurs. The contract does not call for any repayment in period 2, since promises to repay are not credible when the firm's liquidation value is zero. Liquidating the firm in period 1 is inefficient, since the firm would earn more than L if it continued to operate. Under these assumptions, the firm's owners always repay in period 1 when the firm is successful, since they benefit from retaining control and collecting R_2 in the following period. But if the firm fails, then its owners default and creditors liquidate it. Thus there is no strategic default, but filtering failure occurs since there is inefficient liquidation. If lenders instead allowed owners to remain in control following default, then there would be no filtering failure but a high level of strategic default. Because of incomplete information, strategic default and filtering failure cannot both be eliminated.

Bankruptcy law also affects managers' choice of how much effort to expend and whether to delay filing for bankruptcy. Povel (1999) analyses a model in which managers make an effort-level decision and also receive an early signal on whether the firm will succeed. When the signal is bad, managers decide whether to file for bankruptcy or continue operating outside of bankruptcy. Filing for bankruptcy is assumed to be economically efficient in this situation, since it allows creditors to rescue the firm. Neither the effort-level decision nor the signal is observed by creditors. Povel considers two different bankruptcy laws: reorganization and liquidation. In the model, if the bankruptcy procedure is reorganization, the result is that managers choose low effort and file for bankruptcy when the signal is bad. Filing for bankruptcy is economically efficient, but low effort by managers is inefficient. Conversely, if the bankruptcy procedure is liquidation, the result is that managers choose high effort and avoid bankruptcy when the signal is bad. This trade-off suggests that the better bankruptcy procedure could be either reorganization or liquidation, depending on parameter values. See Berkovitch et al. (1998) for a similar model that explores the efficiency of auctions as an alternative bankruptcy procedure.

There is a large literature on reforms of bankruptcy law. Most studies start from the premise that too many firms reorganize in bankruptcy under current law, since reorganization under Chap. 11 has both high transactions costs and high costs of filtering failure. One proposal is to auction all bankrupt firms and use the proceeds to repay creditors according to the APR. This procedure has the dual advantages that it would be quick and that the new owners would make efficient decisions on whether to save or liquidate each firm (see Baird 1986). Another proposal is to use options to divide the value of firms in reorganization (Bebchuk 1988). Both auctions and options would establish a market value of the firm's assets, so that creditors could be repaid according to the APR and deviations from the APR could be eliminated. Another proposal, called bankruptcy contracting, would allow debtors and creditors to adopt their own bankruptcy procedure when they write their loan contracts, rather than requiring them to use the state-supplied mandatory bankruptcy procedure. Schwartz (1997) showed that bankruptcy contracting could improve efficiency in particular circumstances. But whether bankruptcy contracting or any of the other reform proposals would work well in a general model that takes account of other complications – such as the existence of multiple creditor groups and strategic default – has not been established.

Empirical Research

Now we turn to empirical research on corporate bankruptcy. It has focused on measuring the costs of bankruptcy and the size and frequency of deviations from the APR. Studies of the costs of bankruptcy include only the legal and administrative costs of the bankruptcy process; that is, the costs of bankruptcy-induced disruptions are excluded. Most studies have found that bankruptcy costs as a fraction of the value of firms'

assets are higher in liquidation than in reorganization, but this may reflect the fact that bankruptcy costs are subject to economies of scale and larger firms tend to reorganize rather than liquidate in bankruptcy. Unsecured creditors generally receive nothing in liquidation, but are repaid one-third to one-half of their claims in reorganization. This higher return in reorganization could be due to selection bias, if firms that reorganize are in relatively better financial condition. Other studies provide evidence that Chap. 11 filings are associated with an increase in managers' and directors' turnover, suggesting that the process is very disruptive. In addition, many firms that reorganize in Chap. 11 end up requiring additional financial restructuring within a short period. This is consistent with the theoretical prediction that too many financially distressed firms reorganize. Deviations from the APR have been found to occur in around three-quarters of all reorganization plans of large corporations in bankruptcy (see Bris et al. 2006, for a recent study and references).

Personal Bankruptcy

When an individual or a married couple files for bankruptcy under Chap. 7 (the most commonly used procedure), most unsecured debts are discharged. Debtors are obliged to use their non-exempt assets to repay debt, but their future earnings are entirely exempt under the 'fresh start'. Exemption levels, unlike other features of US bankruptcy law, differ across states. The most important exemption is the 'homestead' exemption for equity in owner-occupied homes, which varies widely from zero to unlimited. Because debtors can convert non-exempt assets such as bank accounts into home equity before filing for bankruptcy, high homestead exemptions protect all types of wealth for debtors who are homeowners.

There is also a second personal bankruptcy procedure, Chap. 13, under which debtors' assets are completely exempt, but they must use some of their future earnings to repay their debt. Until recently, debtors had the right to choose between the two procedures and, since most debtors have few non-exempt assets, Chap. 7 was almost always the more favourable. It was also the more heavily used – about 70 per cent of all personal bankruptcy filings were under Chap. 7. Those debtors who filed under Chap. 13 often repaid only token amounts, since the value of their non-exempt assets was zero. However, in late 2005 bankruptcy reforms went into effect that will force some debtors having higher incomes to file for bankruptcy under Chap. 13 and to repay more.

Theory

From an economic standpoint, the main reason for having a personal bankruptcy procedure is to provide individual debtors with consumption insurance by discharging debt when the obligation to repay would cause a substantial reduction in their consumption levels. This is because sharp falls in consumption can have permanent negative effects – debtors may become homeless, their illnesses may become disabilities for lack of medical care, and their children may leave school prematurely and have lower future earnings. Consumption insurance is mainly provided by the public sector in the form of the social safety net – welfare payments, food stamps and health insurance for the poor. But bankruptcy reduces the cost to the public sector of providing the safety net, since discharge of debt in bankruptcy frees up funds for consumption that debtors might otherwise use to repay debt.

The higher the exemption levels for wealth and earnings in bankruptcy, the more the consumption insurance that bankruptcy provides. Theoretical research on personal bankruptcy has focused on deriving optimal exemption levels. Higher levels of both exemptions benefit debtors by providing them with extra consumption insurance, but harm those who repay their debts by reducing the availability of credit and increasing interest rates. However, the two exemptions have differing effects on debtors' incentives to work after bankruptcy. A higher wealth exemption is likely to have little effect on work incentives, while a higher earnings exemption increases debtors' incentive to work as long as the positive substitution effect outweighs the negative income effect.

The model suggests that the optimal earnings exemption is 100 per cent – that is, the 'fresh start' – while the optimal wealth exemption is an intermediate level. This is because a higher earnings exemption both encourages debtors to work more after bankruptcy and provides better consumption insurance than a higher wealth exemption. See White (2005).

An important feature of personal bankruptcy law is that it encourages opportunistic behaviour by debtors. Although bankruptcy debt relief is intended for debtors whose consumption has fallen sharply due to factors such as job loss or illness, in fact debtors' incentive to file is hardly affected by these adverse events. Debtors' financial benefit from bankruptcy equals the amount of debt discharged minus the sum of non-exempt assets that must be used to repay plus the costs of bankruptcy. White (1998b) calculated that at least one-sixth of US households would benefit financially from filing for bankruptcy, and this figure rose to more than one- half if households were assumed to pursue various strategies, such borrowing more on an unsecured basis, converting non-exempt assets into exempt home equity, and moving to states with high homestead exemptions. White (1998b) also found that these calculations understate the proportion of households that would benefit from bankruptcy, since some households that would not benefit from filing immediately could benefit from filing in the future. She calculated the value of the option to file for bankruptcy and found that it is particularly valuable for high-wealth households and those in high-exemption states. These features of bankruptcy law are probably responsible for high filing levels (more than 1.6 million US households filed for bankruptcy in 2003) and for the fact that the US Congress recently changed Chap. 7 to make bankruptcy less attractive to many debtors.

Empirical Research

Most of the empirical research on personal bankruptcy makes use of the variation in exemption levels that causes bankruptcy law to differ across US states. Gropp et al. (1997) found that, if households live in states with high rather than low exemptions, they are more likely to be turned down for credit, they borrow less, and they pay higher interest rates. They also found that in high-exemption states credit is redistributed from low-asset to high-asset households. Households in high- exemption states demand more credit because borrowing is less risky, but lenders respond by offering larger loans to high-asset households while rationing credit more tightly to low-asset households. Fay, Hurst and White (2002) found that households are more likely to file for bankruptcy when their financial benefit from filing is higher. Since households' financial benefit from filing is positively related to the size of the exemption, this means that households are more likely to file if they live in states with higher bankruptcy exemptions. Fay, Hurst and White did not find that recent job loss or health problems were significantly related to whether households filed for bankruptcy. But they found that households were more likely to file when they live in regions that have higher average bankruptcy filing rates – which suggests the existence of network effects.

Personal bankruptcy exemption levels also affect small businesses, since business debts often are personal obligations of the business owner and these debts are discharged in bankruptcy. Fan and White (2003) found that individuals are more likely to own or start businesses in states with higher exemption levels, presumably because the additional consumption insurance in these states makes going into business more attractive by lowering the cost of failure. But Berkowitz and White (2004) found that small businesses are more likely to be turned down for credit and to pay higher interest rates if they are located in states with higher exemption levels. Overall, higher exemption levels have mixed effects on small business.

Finally, since higher exemption levels provide households with additional consumption insurance, the variance of household consumption is predicted to be smaller in states that have higher exemption levels. Grant (2006) found macro-level support for this hypothesis using data on the variance of consumption across state – years.

See Also

► Bankruptcy, Economics Of

Bibliography

Baird, D. 1986. The uneasy case for corporate reorganizations. *Journal of Legal Studies* 15: 127–147.
Bebchuk, L. 1988. A new method for corporate reorganization. *Harvard Law Review* 101: 775–804.
Bebchuk, L. 2002. The ex ante costs of violating absolute priority in bankruptcy. *Journal of Finance* 57: 445–460.
Berkovitch, E., R. Israel, and J. Zender. 1998. The design of bankruptcy law: A case for management bias in bankruptcy reorganizations. *Journal of Financial and Quantitative Analysis* 33: 441–467.
Berkowitz, J., and M. White. 2004. Bankruptcy and small firms' access to credit. *RAND Journal of Economics* 35: 69–84.
Bolton, P., and D. Scharfstein. 1996. Optimal debt structure and the number of creditors. *Journal of Political Economy* 104: 1–25.
Bris, A., I. Welch, and N Zhu. 2006. The costs of bankruptcy: Chapter 7 cash auctions vs. Chapter 11 bargaining. *Journal of Finance* 61.
Fan, W., and M. White. 2003. Personal bankruptcy and the level of entrepreneurial activity. *Journal of Law and Economics* 46: 543–568.
Fay, S., E. Hurst, and M. White. 2002. The household bankruptcy decision. *American Economic Review* 92: 706–718.
Grant, C. 2006. Evidence on the effect of U.S. bankruptcy exemptions. In *Economics of consumer credit: European experience and lessons from the U.S.*, ed. G. Bertola, R. Disney, and C. Grant. Cambridge, MA: MIT Press.
Gropp, R., K. Scholz, and M. White. 1997. Personal bankruptcy and credit supply and demand. *Quarterly Journal of Economics* 112: 217–252.
Povel, P. 1999. Optimal soft or tough bankruptcy procedures. *Journal of Law, Economics, and Organization* 15: 659–684.
Schwartz, A. 1997. Contracting about bankruptcy. *Journal of Law, Economics, and Organization* 13: 127–146.
White, M. 1998a. Why don't more households file for bankruptcy? *Journal of Law, Economics, and Organization* 14: 205–231.
White, M. 1998b. Why it pays to file for bankruptcy: A critical look at incentives under U.S. bankruptcy laws and a proposal for change. *University of Chicago Law Review* 65: 685–732.
White, M. 2002. Sovereigns in distress: Do they need bankruptcy? *Brookings Papers on Economic Activity* 2002(1): 287–319.
White, M. 2005. *Personal bankruptcy: Insurance, work effort, opportunism and the efficiency of the 'Fresh Start'*. Paper presented to the American Law and Economics Association conference, New York, May.
White, M. 2006. Bankruptcy law. In *Handbook of law and economics*, ed. A. Mitchell Polinsky and S. Shavell. Amsterdam: North-Holland.

Bankruptcy, Economics of

Arturo Bris

Abstract

Bankruptcy is the formal procedure to resolve the disputes among creditors, shareholders, and managers of a company in financial distress. Countries have designed bankruptcy procedures that differ in the control that is given to the existing management relative to creditors. These differences determine the incentives that are given to the parties before, during, and after the bankruptcy proceedings. They also determine how expensive the bankruptcy process is. Ultimately, bankruptcy costs are borne by firms' shareholders.

Keywords

Absolute priority rule (APR); Administrative receiverships; Bankruptcy; Bankruptcy law; Bankruptcy, economics of; Chapter 11 reorganizations; Chapter 7 liquidations (USA); Default; Liquidations; Workouts

JEL Classification

K3; K35

Bankruptcy is the legal procedure whereby the assets of a debtor are distributed among its creditors. The debtor can be either an individual or a firm. In corporations, bankruptcy happens when either the firm or its creditors delegate a third party – be it a judge or other public official – to determine the amount of the creditors' claims, as well as the way to distribute the firm's assets among them. In essence, bankruptcy results from financial distress, which happens when the

market value of the assets is insufficient to satisfy the debt claims, or when the firm does not generate enough cash flow to meet the coupon and interest payments. An alternative to bankruptcy is an informal reorganization, or *workout,* whereby creditors relax debt covenants, possibly exchanging their claims for a package of new claims.

Bankruptcy is an old European institution that derives its name from the Italian '*banca rotta'* (broken bench). It refers to the boards from which traders in medieval towns traded coins, and which they broke whenever they defaulted on their payments. Nowadays, countries have implemented different procedures to deal with the distribution of the assets of a firm that cannot meet its debt obligations. In the United States, firms and creditors can opt into two forms of restructuring. Under a Chapter 7 liquidation, assets are sold piecemeal and the proceeds distributed according to the *absolute priority rule* (APR), whereby debt and equity are paid according to a predetermined order: secured debt first, then unsecured claims, and finally common stock. The distinction between *senior* and *junior* claims refers to the priority of secured debt (senior) over unsecured debt (junior). The firm ceases to exist after a Chapter 7. Under a Chapter 11 reorganization, shareholders and creditors agree on a reorganization plan, which allows the company to continue. When the company enters a Chapter 11, the firm becomes a 'debtor-in-possession', a term that recognizes that the management retains control of the company's operations, under court supervision. In a Chapter 11, APR may be violated if secured creditors give up part of their claims in favour of unsecured debtors, or if shareholders receive some interest in the restructured firm at the expense of debtholders (Herbert 1998).

Under the absolute priority rule, unsecured claims are classified into priority claims and general unsecured claims. Priority claims are further classified into three groups: administrative claims, wages and employee benefits, and taxes. This means that, under APR – which is always upheld in Chapter 7 cases – wages cannot be paid unless administrative expenses (compensation of lawyers and other professionals) have been satisfied in full. Moreover, tax claims include only those taxes that the firm owes at the time it files for bankruptcy.

The practice in the United States is to reimburse administrative expenses incurred by the committee of unsecured creditors. A Chapter 11 creditors' committee is composed of creditors 'that hold the seven largest claims against the debtor of the kinds represented on such committee' (Bankruptcy Code §1102(b)(1)). The bankruptcy court is authorized to reimburse a substantial portion of the expert expenses that juniors incur. However, the United States code does not authorize the bankruptcy court to compensate the expenses of creditors whom it defines as 'senior.' This cost allocation fails to encourage the seniors to spend on activities that increase the value of the firm, but encourages the juniors to spend on activities that maximize only the value of their own claims.

In the United States the debtor has an exclusivity period of 120 days to file a plan of reorganization. This period can be, and usually is, extended upon the debtor's requests. In the plan, each class of creditors is classified as *impaired* or *unimpaired.* An unimpaired class of creditors is paid in full, and does not vote on the reorganization plan. The plan requires the approval of each impaired class of creditors and equity security holders. Approval requires dual majority: more than one-half of the votes, and more than two-thirds of the amount of the claims.

In the United Kingdom and other countries with British legal traditions, such as Canada, Australia and New Zealand, bankrupt companies are restructured via an *administrative receivership.* White (1996) and Franks and Davydenko (2006) provide a comparison between the bankruptcy codes in the United States and some European countries. Under an administrative receivership, the secured creditors appoint an expert (the administrative receiver) whose objective is to obtain sufficient funds to repay the secured creditors. To do that, the receiver can either liquidate some assets or sell the company as a going

concern. The receiver does not have any obligation with respect to other creditors or shareholders, as long as absolute priority is respected. Unlike with a United States Chapter 11, in a receivership control is transferred from the management to the secured creditors.

Under the old French system neither the firm nor the creditors retained control. The court appointed an administrator who managed the day-to-day operations of the firm, and whose objectives were, first, to preserve the estate and employment, and then to satisfy creditors. Most systems in Continental Europe have followed this tradition. In the new Loi de Sauvegarde des Enterprises enacted in 2005, France has moved towards the Chapter 11 system in the United States.

In Germany, the system introduced in 1999 establishes an automatic stay of three months, which means that creditors cannot dispose of the firm's assets during that period. Moreover, and similar to a Chapter 7 in the United States, the court appoints an administrator who monitors the process and determines a plan of reorganization.

Auctions are a very efficient alternative to court-administered procedures. In Sweden, the court appoints an independent trustee who is in charge of selling the firm's assets to the highest bidder. The winning bidder can pay only in cash, as described in Thorburn (2000), and the trustee distributes the proceeds respecting the AP-R. Stromberg (2000) shows that in one out of three cases in Sweden the assets are sold back to the incumbent managers (because they have the highest valuation of the assets), and the remaining cases are liquidated.

Controversy Over Chapter 11

In recent years, there has been a convergence in bankruptcy laws towards a Chapter 11-type reorganization. Countries in western and eastern Europe, Asia and Latin America have enacted regulations that allow managers to retain control of defaulted firms. Regulators have moved from a system that favours liquidations to a legal procedure that tends to maximize the probability of firm survival. However, the efficiency of Chapter 11 has been questioned by scholars like Bebchuk (1988), Adler (1993), Schwartz (1998), Baird and Rasmussen (2002), and Baird and Morrison (2005). They promote a *contractual approach* to bankruptcy, or a formal scheme of bargained bankruptcy. Under this view, the parties should be free to bargain in advance over a set of rules that will govern their rights in the event of bankruptcy, with Chapter 11 being only a default system. Bebchuk (1988), for instance, proposes that firms can issue derivative securities, contingent on the firm being in default. The contractual view attacks the Chapter 11 system on several fronts, first of all on the grounds that it leads to inefficient outcomes (Baird and Morrison 2005; Franks and Loranth 2006). In particular, Franks and Loranth show that Chapter 11 in Hungary is biased in favour of inefficient going concerns. The argument is that most bankrupt firms should be liquidated rather than reorganized. Chapter 11 is also attacked because it is considered a more lengthy process than other systems (Stromberg 2000; Thorburn 2000). Additionally, it is extremely expensive (Bris et al. 2006).

The opponents of such a private bankruptcy system (Warren and Westbrook 2005) make two important arguments to defend Chapter 11. In principle, a private system would have only redistributive effects, with some creditors (secured and large creditors) shifting risks to others. Also, Chapter 11 is a mechanism by which benevolent large creditors give up part of their claims in favour of small, empowered creditors. Therefore it has a positive redistributive effect. Finally, a private system is inefficient because of the duplication of transaction costs.

Most of the theoretical and empirical research on bankruptcy addresses the conflicts that arise among creditors, shareholders, firm managers and bankruptcy specialists. These conflicts arise during the bankruptcy proceedings, but also when the company is in financial distress and before it files for bankruptcy. The design of the bankruptcy system can affect the interaction among all these agents, the efficiency of the bankruptcy process and, therefore, the costs of bankruptcy.

Incentives Before Filing for Bankruptcy

Financial distress may lead to bankruptcy if either the firm management or the creditors opt into a legal procedure to resolve their disputes. But, if the distressed firm is economically viable, managers have an incentive to delay filing for bankruptcy and to maintain operations, especially if the legal procedure gives control to a third party. Self-interested managers will then preserve their jobs at the expense of shareholders and creditors. Jensen and Meckling (1976) show that in distressed firms there is a *debt overhang* problem. Managers have an incentive to bypass positive net present value (NPV) projects (a problem known as *underinvestment*) because they benefit only current creditors (Myers 1977). Instead, when choosing between less and more risky projects managers prefer to invest in more risky projects because managers act on behalf of shareholders, and shareholders, because of limited liability, are interested only in the upside of the investments *(excess risk taking* or *overinvestment).* These incentives in turn reduce the value of the debtor's claims and ultimately the value of the firm because creditors take them into account when pricing their securities.

Recently, Adler et al. (2005) have shown that a change in regulation in the United States around 2000, which gave more control to creditors during the filing period, induced managers to delay the bankruptcy filing. Indeed, they show that after 2000 firms that file for Chapter 11 in the United States display a worse financial and operating condition. This can explain why, in countries with secured creditor control of the bankruptcy process, the number of bankruptcy filings is much lower, and firm managers prefer liquidation (Claessens and Kappler 2005).

Conversely, and depending on the debt structure, managers may have an incentive to default strategically even if the firm is still economically viable. Bolton and Scharfstein (1996) argue that managers will always prefer to default strategically so as to divert cash to themselves. In order to avoid that distortion, creditors should have the right to liquidate the firm in case of default. However, this induces inefficient liquidations because the value of the firm as a going concern may exceed its liquidation value. Bolton and Scharfstein (1996) show that borrowing from multiple creditors solves the problem by increasing the liquidation value of the firm.

Incentives During Bankruptcy Proceedings

The efficiency of the bankruptcy process and a firm's capital structure are closely related because, for a firm with multiple creditors, bankruptcy results in coordination problems among creditors, as well as conflicts between secured and unsecured, or between senior and junior, claimants. Regarding coordination problems, and in contrast to Bolton and Scharfstein (1996), Bris and Welch (2005) argue that, when competing for the firm's assets, multiple creditors (similar to public bonds) waste the firm's resources in fighting with each other; hence, it is more efficient to issue highly concentrated debt (bank debt). Indeed, Welch (1997) shows that bank debt should be senior because a single creditor fights better with shareholders, thereby increasing the *ex ante* value of the debt.

Conflicts between secured and unsecured creditors depend on the bankruptcy system and the priority rules. If unsecured creditors can extract rents at the expense of more senior debtors (that is, if absolute priority can be violated), then a firm may prefer to liquidate its assets because unsecured creditors will expend the firm's resources in order to satisfy part of their claim. Eberhart et al. (1990) and Franks and Torous (1994) show that APR is often violated under Chapter 11.

Firms in bankruptcy are allowed sometimes to issue new financing that can be senior to the already outstanding debt (*debtor-in-possession, DIP, financing*). The ability to raise DIP financing is priced *ex ante* by the firm's creditors. Therefore, it increases the value of the firm *ex post* but it reduces shareholder value *ex ante.* This trade-off has been extensively considered in the literature.

Life After Bankruptcy

The design of the bankruptcy process can also affect the performance of firms when they emerge from Chapter 11. Hotchkiss (1995) reports that over 40% of the firms in her sample still experience operating losses in the three years following the bankruptcy case, while another 32% re-file for bankruptcy or restructure their debt.

Bankruptcy Costs

Bankruptcy costs encompass not only the explicit payments made to bankruptcy specialists (lawyers, trustees, accountants, investment bankers) but also the indirect costs of being in default. Among the latter, we can include loss of customers when the company is in financial distress, adverse payment terms enforced by suppliers when the viability of the firm is not guaranteed, loss of key personnel and waste of management time.

Measuring the indirect costs of bankruptcy is very difficult. Altman (1984) uses forgone profits as a proxy, while Opler and Titman (1994) focus on losses of trade credit. However, because of the nature of the indirect costs, any proxy tends to underestimate their extent. Other researchers have used the length of the proceedings as a proxy for indirect bankruptcy costs, under the assumption that, the longer the firm stays in bankruptcy, the larger the collateral effects (Franks and Torous 1994). Bris et al. (2006) show that both liquidations under Chapter 7 and reorganizations under Chapter 11 take about two years to resolve. In exploring the Swedish system, Thorburn (2000) shows that the Swedish auction system is much faster than the United States Chapter 11 process, since auctions take only two months on average.

The evidence on direct costs is more extensive. Warner (1977) finds that the direct costs of bankruptcy are about 4% of the market value of the firm one year prior to the default. This result is based on a sample of 11 bankrupt railroads. Altman (1984) calculates these costs to be about 7.5% of firm value, using a broader sample of 19 bankrupt companies from 1974 to 1978. Using 105 Chapter 11 cases, Ang et al. (1982) report that administrative fees are about 7.5% of the total liquidating value of the bankrupt corporation's assets. Lubben (2000) calculates in his sample of 22 firms from 1994 that the cost of legal counsel in Chapter 11 bankruptcy represents 1.8% of the distressed firm's total assets, and in some cases more than 5%. In his average case, the debtor spends $500,000 on lawyers, and creditors spend $230,000. LoPucki and Doherty (2004) study a sample of 48 cases from 1998 to 2002, mostly from Delaware and New York. They report that professional fees were 1.4% of the debtors' total assets at the beginning of the bankruptcy case. Bris et al. (2006) compare the costs of bankruptcy for Chapter 7 and Chapter 11 cases. They report that the mean ratio of total expenses to assets is 9.5% for Chapter 11, and 8.1% for Chapter 7. However, they warn against simple averages because cost measures depend on the value of the assets (pre-bankruptcy or post-bankruptcy) one uses.

Conclusion

The design of a bankruptcy system is very important because it determines shareholder value for all firms, whether or not they are in financial distress. The reason is that any conflict that can arise among creditors of different classes, and any coordination problem in the bankruptcy proceedings among creditors in a similar class, are both priced in the debt securities that a company issues. Moreover, the bankruptcy system can impose distortions on a firm's policies when it is in financial distress; in particular it can induce managers to make suboptimal decisions at the expense of shareholders.

Countries' legal systems differ in terms of who controls the firm's assets during bankruptcy. Because control shapes the conflicts set out above, this feature of the bankruptcy system is one of the most important considered by the academic literature. Additionally, scholars have studied the issue of bankruptcy costs in detail. While we have extensive evidence on the direct costs of bankruptcy, the indirect costs of being in distress are very difficult to measure.

See Also

- ▶ Bankruptcy Law, Economics of Corporate and Personal
- ▶ Default and Enforcement Constraints
- ▶ Extremal Quantiles and Value-at-Risk

Bibliography

Adler, B. 1993. Financial and political theories of American corporate bankruptcy. *Stanford Law Review* 45: 311–346.

Adler, B., V. Capkun, and L. Weiss. 2005. *Theory and evidence on the bankruptcy initiation problem*. Working paper, University of Lausanne.

Altman, E. 1984. A further empirical investigation of the bankruptcy cost question. *Journal of Finance,* 1067–1089

Ang, J., J. Chua, and J. McConnell. 1982. The administrative costs of corporate bankruptcy: A note. *Journal of Finance* 37: 219–226.

Baird, D., and E. Morrison. 2005. Serial entrepreneurs and small business bankruptcies. *Columbia Law Review* 105: 2310–2368.

Baird, D., and R. Rasmussen. 2002. The end of bankruptcy. *Stanford Law Review* 55: 751–790.

Bebchuk, L. 1988. A new approach to corporate reorganizations. *Harvard Law Review* 101: 775–804.

Bolton, P., and D. Scharfstein. 1996. Optimal debt structure and the number of creditors. *Journal of Political Economy* 104: 1–25.

Bris, A., and I. Welch. 2005. The optimal concentration of creditors. *Journal of Finance* 60: 2193–2212.

Bris, A., I. Welch, and N. Zhu. 2006. The costs of bankruptcy: Chapter 7 liquidations vs. Chapter 11 reorganizations. *Journal of Finance* 61: 1253–1303.

Claessens, S., and L. Kappler. 2005. Bankruptcy around the world: Explanations of its relative use. *American Law and Economics Review* 7: 253–283.

Eberhart, A., W. Moore, and R. Roenfeldt. 1990. Security pricing and deviations from absolute priority rule in bankruptcy proceedings. *Journal of Finance* 45: 1457–1469.

Franks, J., and S. Davydenko. 2006. *Do bankruptcy codes matter? A study of defaults in France, Germany and the UK*, Finance Working Paper No. 89/2005. Brussels: European Corporate Governance Institute.

Franks, J. and G. Loranth. 2006. *A study of inefficient going concerns in bankruptcy*. Working paper, London Business School.

Franks, J., and W. Torous. 1994. A comparison of financial recontracting in distressed exchanges and Chapter 11 reorganizations. *Journal of Financial Economics* 35: 349–370.

Herbert, M. 1998. *Understanding Bankruptcy*. New York: Matthew Bender & Company.

Hotchkiss, E. 1995. Post-bankruptcy performance and management turnover. *Journal of Finance* 50: 3–22.

Jensen, M., and W. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

LoPucki, L., and J. Doherty. 2004. The determinants of professional fees in large bankruptcy reorganization cases. *Journal of Empirical Legal Studies* 1: 111–141.

Lubben, S. 2000. The direct costs of corporate reorganization: An empirical examination of professional fees in large Chapter 11 cases. *American Bankruptcy Law Journal* 74: 508–552.

Myers, S. 1977. The determinants of corporate borrowing. *Journal of Financial Economics* 5: 147–175.

Opler, T., and S. Titman. 1994. Financial distress and corporate performance. *Journal of Finance* 49: 1015–1040.

Schwartz, A. 1998. A contract theory approach to business bankruptcy. *Yale Law Journal* 107: 1807–1852.

Stromberg, P. 2000. Conflicts of interest and market illiquidity in bankruptcy auctions: Theory and tests. *Journal of Finance* 55: 2641–2692.

Thorburn, K. 2000. Bankruptcy auctions: Costs, debt recovery and firm survival. *Journal of Financial Economics* 58: 337–368.

Warner, J. 1977. Bankruptcy cost: Some evidence. *Journal of Finance* 32: 337–347.

Warren, E., and J. Westbrook. 2005. Contracting out of bankruptcy: An empirical intervention. *Harvard Law Review* 118: 1197–1224.

Welch, I. 1997. Why is bank debt senior? A theory of asymmetry and claim priority based on influence costs. *Review of Financial Studies* 10: 1203–1236.

White, M. 1996. The costs of corporate bankruptcy: A U.S.–European comparison. In *Corporate bankruptcy: Economic and legal perspectives*, ed. J. Bhandari and L. Weiss. Cambridge: Cambridge University Press.

Banks, Jeffrey Scot (1958–2000)

David Austen-Smith

Keywords

Adverse selection; Agency models; Banks Set; Banks, J; Divinity; Incomplete information; Industrial organization; Moral hazard; Nash equilibrium; Pretrial bargaining; Social choice; Spatial theory of elections; Subgame perfection

JEL Classification
B31

Jeff Banks received his BA from University of California, Los Angeles, in 1982 and his Ph.D. from California Institute of Technology in 1986. He arrived as a new assistant professor of political science and economics at the University of Rochester with two significant and influential publications in hand, reflecting his principal interests in social choice theory (1985) and game theory (1987) respectively. By the time he died of complications from treating leukemia, Banks had published (or had forthcoming) more than 50 papers in economics, game theory and formal political theory, edited one conference volume, published a review monograph and coauthored two books.

In the 1985 paper, Banks completely characterized the set of subgame perfect Nash equilibrium outcomes achievable through an amendment agenda on a voting tournament. In effect, this set (which came to be called the Banks Set through no fault of its author) defines the consequential limits of an agenda-setter's power under the amendment procedure. Banks went on to write a series of influential papers on a variety of topics in social choice theory (for example, 1995; 1996; 2000; 2006) and in more applied positive political theory (for example, 1988; 1989; 1990a; 1990b). Indeed, it is difficult to identify any area within the field to which Banks did not make some significant contribution.

In (1987), Banks addressed the equilibrium refinement problem. Their proposed refinement, 'divinity', is on out-of-equilibrium beliefs and is closely related to the Cho and Kreps (1987) D1 refinement. Like D1, a virtue of divinity (in particular of its stronger variant, universal divinity) is that it is widely applicable and easy to compute, especially in games with a continuum of types and actions. Banks was a pioneer in developing strategic theories of collective decision-making under incomplete information, and his (1990a) paper is both the seminal contribution to the spatial theory of elections under incomplete information and the first application of divinity to an applied problem. Subsequently, the refinement has been used profitably by others on a variety of problems in industrial organization, pretrial bargaining and so forth. Along with incomplete information, Banks contributed some of the earliest formal papers dealing with problems of time and dynamics in politics. For example, he explored dynamic agency models that exhibit both moral hazard and adverse selection simultaneously (1993; 1998). Such environments are notoriously complicated and, as a step towards developing an appropriate toolbox for handling them, Banks (1992) made an important contribution to theory of denumerably armed bandits.

Banks's professional career barely spanned 15 years, yet the footprint he has left on (especially) positive political theory is considerable. He was a fine teacher and a remarkable colleague; he is, and will continue to be, much missed.

Selected Works

1985. Sophisticated voting outcomes and agenda control. *Social Choice and Welfare* 1: 295–306.
1987. (With J. Sobel.) Equilibrium selection in signaling games. *Econometrica* 55: 647–661.
1988. (With D. Austen-Smith.) Elections, coalitions and legislative outcomes. *American Political Science Review* 82: 405–422.
1989. Agency budgets, cost information and auditing. *American Journal of Political Science* 33: 670–699.
1990a. A model of electoral competition with incomplete information. *Journal of Economic Theory* 50: 309–325.
1990b. Monopoly agenda control and asymmetric information. *Quarterly Journal of Economics* 105: 445–464.
1992. (With R. Sundaram.) Denumerable-armed bandits. *Econometrica* 60: 1071–1096.
1993. (With R. Sundaram.) Adverse selection and moral hazard in a repeated elections model. In *Political economy: Institutions, information and competition*, ed. W. Barnett, M. Hinich and N. Schofield. New York: Cambridge University Press.
1995. Acyclic social choice from finite sets. *Social Choice and Welfare* 12: 293–310.

1996. (With D. Austen-Smith.) Information aggregation, rationality and the Condorcet jury theorem. *American Political Science Review* 90: 405–422.
1998. (With R. Sundaram.) Optimal retention in agency problems. *Journal of Economic Theory* 82: 293–323.
2000. (With J. Duggan.) A bargaining model of collective choice. *American Political Science Review* 94: 73–88.
2006. (With J. Duggan.) A multidimensional model of repeated elections. *Quarterly Journal of Political Science* 1: 49–85.

Bibliography

Cho, I., and D. Kreps. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102: 179–221.

Baran, Paul Alexander (1910–1964)

Paul M. Sweezy

Keywords

Baran, P. A.; Capitalism; Feudalism; India, economics in; Japan, economics in; Monopoly capitalism; Potential surplus; Surplus; Surplus value; Sweezy, P. M.; Underdevelopment

JEL Classifications

B31

Paul Baran, the eminent Marxist economist, was born on 8 December 1910 in Nikolaev, Russia, the son of a medical doctor who was a member of the Menshevik branch of the Russian revolutionary movement. After the October Revolution the family moved to Germany, where Baran's formal education began. In 1925 the father was offered a position in Moscow and returned to the USSR. Baran began his studies in economics at the University of Moscow the following year. Both his ideas and his politics were deeply and permanently influenced by the intense debates and struggles within the Communist Party in the late 1920s. Offered a research assignment at the Agricultural Academy in Berlin in late 1928, he enrolled in the University of Berlin, and when his assignment at the Agricultural Academy ended he accepted an assistantship at the famous Institute for Social Research in Frankfurt. This experience too had a lasting influence on his intellectual development.

Leaving Germany shortly after Hitler's rise to power, Baran sought without success to find academic employment in France. He therefore moved to Warsaw, where his paternal uncles had a flourishing international lumber business. During the next few years he travelled widely as a representive of his uncles' business, ending up in London in 1938. With the approach of World War II, however, he decided to take what savings he had been able to accumulate, move to the United States, and resume his interrupted academic career.

Arriving in the United States in the fall of 1939, he was accepted as a graduate student in economics at Harvard. From there he went to wartime Washington, where he served in the Office of Price Administration, the Research and Development branch of the Office of Strategic Services, and the United States Strategic Bombing Survey, ending in 1945–6 as Deputy Chief of the Survey's mission to Japan. Back in the United States, he took a job at the Department of Commerce and gave lectures at George Washington University before being offered a position in the Research Department of the Federal Reserve Bank of New York. After three years in New York, he accepted an offer to join the economics faculty at Stanford University and was promoted to a full professorship in 1951, a position he retained until his death of a heart attack on 26 March 1964.

Baran was not a prolific writer, but his two main books, *The Political Economy of Growth* (1957) and (in collaboration with Paul M. Sweezy) *Monopoly Capital*: *An Essay on the American Economic and Social Order* (1966), are generally considered to be among the most

important works in the Marxian tradition of the post-World War II period.

The Political Economy of Growth is concerned with the processes and condition of economic growth (or development, the terms are used interchangeably) in both industrialized and underdeveloped societies, with a special emphasis throughout on the ways the two relate to and interact with each other. It is at once an outstanding work of scholarship weaving an intricate pattern of theory and history, and a passionate polemic against mainstream economics. Its chief (innovative) analytical concept is that of 'potential surplus', defined as 'the difference between the output that *could* be produced in a given natural and technological environment with the help of employable productive resources, and what might be regarded as essential consumption'. (This concept presupposes Marx's 'surplus value', extending and modifying it for the particular purposes of the study in hand.) Two long chapters, totalling 90 pages, apply the concepts of surplus and potential surplus to the analysis of monopoly capitalism in ways that would later be refined and elaborated in *Monopoly Capital.* Three chapters (115 pages) follow on 'backwardness' (also called underdevelopment), and it is for these that the book has become famous, especially in the Third World.

Baran begins this analysis with a question which may be said to define the focus of the whole work: 'Why is it that in the backward capitalist countries there has been no advance along the lines of capitalist development that are familiar from the history of other capitalist countries, and why is it that forward movement there has been slow or altogether absent?' His answer, in briefest summary, is as follows: all present-day capitalist societies evolved from precapitalist conditions which Baran for convenience labels 'feudal' (explicitly recognizing that a variety of social formations are subsumed under this heading). Viable capitalist societies could have emerged in various parts of the world; actually the decisive breakthrough occurred in Western Europe (Baran speculates on the reasons, but in any case they are not crucial to the subsequent history). Having achieved its headstart, Europe proceeded to conquer weaker precapitalist countries, plunder their accumulated stores of wealth, subject them to unequal trading relations, and reorganize their economic structures to serve the needs of the Europeans. This was the origin of the great divide in the world capitalist system between the developed and the underdeveloped parts. As the system spread into the four corners of the globe, new areas were added, mostly to the underdeveloped part but in a few cases to the developed (North America, Australia, Japan). One of the highlights of Baran's study is the brilliant historical sketch of the contrasting ways India and Japan were incorporated into the world capitalist system, the one as a hapless dependency, the other as a strong contender for a place at the top of the pyramid of power. Baran's message to the Third World was loud and clear: once trapped in the world capitalist system, there is no hope for genuine progress; only a revolutionary break can open the road to a better future. The message has been widely heard. Most of the revolutionary movements of the Third World have been deeply influenced, directly or indirectly, by Paul Baran's *Political Economy of Growth.*

The economic analysis of *Monopoly Capital* is a development and systematization of ideas already contained in *the Political Economy of Growth* and Paul Sweezy's *The Theory of Capitalist Development* (1942). The central theme is that in a mature capitalist economy dominated by a handful of giant corporations the potential for capital accumulation far exceeds the profitable investment opportunities provided by the normal *modus operandi* of the private enterprise system. This results in a deepening tendency to stagnation which, if the system is to survive, must be continuously and increasingly counteracted by internal and external factors. In the authors' estimation – not always shared, or even understood by critics – the new and original contributions of *Monopoly Capital* had to do mainly with these counteracting factors and their far-reaching consequences for the history, politics, and culture of American society during the period from roughly the 1890s to the 1950s when the book was written. They intended it, in other words, as much more than a work of economics in the usual meaning of the terms.

See Also

▶ Monopoly Capitalism

Selected Works

There is a comprehensive bibliography of Baran's writings in English in a special issue of *Monthly Review*, 'In Memory of Paul Alexander Baran. Born at Nikolaev, the Ukraine, 8 December 1910. Died at San Francisco, California, 26 March 1964', 16(11), March 1965. This also includes statements on his life and work by more than three dozen contributors, most of whom had been his friends or colleagues.

1957. *The political economy of growth.* New York: Monthly Review Press. 2nd ed, with a new preface, 1962.

1966. (With P.M. Sweezy.) *Monopoly capital: An essay on the American economic and social order*. New York: Monthly Review Press.

1970. *The longer view*: *Essays toward a critique of political economy.* Edited by J. O'Neill, preface by P.M. Sweezy. New York: Monthly Review Press. This volume, which follows an outline prepared before his death by the author, brings together his most important hitherto scattered essays and reviews.

Barbon, Nicholas (1637/40–?1698)

Douglas Vickers

Keywords

Barbon, N.; Child, J.; Money supply; North, D.; Use value

JEL Classifications

B31

Nicholas Barbon, son of Praisegod Barbon, a London leather merchant, was born in 1637 (or 1640), and after studying medicine at Leyden and Utrecht and taking the MD at Utrecht in 1661, was admitted an Honorary Fellow of the College of Physicians at London in 1664. He was elected a Member of Parliament in 1690 and 1695. His successful career in various mercantile activities is reported in the autobiography of Roger North, the brother, biographer, and co-author of Sir Dudley North. He was engaged in the building trade in London following the great fire of 1666, and in 1685 he published a pamphlet *Apology for the Builder*: *or a Discourse showing the Cause and Effects of the Increase of Building.* In 1681 he established the first fire insurance company, and in 1684 published an *Account of two insurance offices.* Barbon also established a large financial venture in banking. With John Asgill he operated a land bank in 1695 and in the same year published *An Account of the Land Bank, showing the design and manner of the settlement*, and prepared a scheme for a national land bank which did not, however, come into existence.

Barbon's place in the history of economics is due to his *Discourse of Trade* (1690) and his more important *Discourse concerning coining the new money lighter*: *An answer to Mr Locke's Considerations about raising the value of money.* Taking the same position as Josiah Child and arguing, against Locke, for a legal reduction of the maximum rate of interest, he published in 1694 An *Answer to ... reasons against reducing interest to four per cent.* His argument against trade restrictions and for international free trade principles places him in the front rank of anticipators of the doctrines that developed in the following century. He exhibited clearly the connection between the supply of money and the effective level of trade. Against the proposals to recoin the currency at the old standard he pointed out the potential deflationary effects of the reduction in the money supply that would result, 'the consequence whereof will be that trade will be at a stand'.

Barbon's concern with the 'disorder... that attends a nation that want money to drive their trade and commerce' and the 'prejudice to the state by making money scarce' led him to argue, in contexts that elevated to priority the functional significance of money, that 'it is not absolutely necessary that

money should be made of gold or silver'. 'Banks of credit... are of great advantage to trade.' 'Money is the instrument and measure of commerce and not silver.' Barbon held a supply and demand theory of market price, based on a logically prior notion of use values, and what he called 'time and place' value. He argued that 'interest is the rent of stock and is the same as the rent of land', claiming that a lower interest rate would raise capital values, indirectly by remedying 'the decay of trade' and directly by increasing the capitalized value of income streams.

Consumption expenditures, Barbon argued, provided employment. In his argument that 'prodigality is a vice that is prejudicial to the man but not to trade ... covetousness is a vice prejudicial to both man and trade', he anticipated the prodigality and employment-creating expenditure argument of the following century.

Selected Works

1690. *A Discourse of Trade.* London, Ed. J.-H. Hollander, Baltimore: Johns Hopkins University Reprint, 1905.
1696. *A discourse concerning coining the new money lighter.* London.

Bibliography

Letwin, W. 1963. *The origins of scientific economics: English economic thought 1660–1776*. London: Methuen.
Vickers, D. 1959. *Studies in the theory of money 1690–1776*. Philadelphia: Chilton.

Bargaining

Roberto Serrano

Abstract

This article is a survey on bargaining theory. The focus is the game theoretic approach to bargaining, both on its axiomatic and strategic counterparts. The application of bargaining theory to large markets and its connections with competitive allocations are also discussed.

Keywords

Asymmetric information; Bargaining; Bayesian equilibrium; Edgeworth, F.; Game theory; Harsanyi, J.; Hicks, J.; Incomplete information; Independence of irrelevant alternatives; Kalai–Smorodinsky bargaining solution; Morgenstern, O.; Nash equilibrium; Nash solution; Nash's demand game; Pareto efficiency; Rational behaviour; Risk aversion; Subgame perfect equilibrium; Ultimatum game; Von Neumann, J.; Walrasian outcomes; Zeuthen, F

JEL Classifications

C7

In its simplest definition, 'bargaining' is a socio-economic phenomenon involving two parties, who can cooperate towards the creation of a commonly desirable surplus, over whose distribution the parties are in conflict.

The nature of the cooperation in the agreement and the relative positions of the two parties in the status quo before agreement takes place will influence the way in which the created surplus is divided. Many social, political and economic interactions of relevance fit this definition: a buyer and a seller trying to transact a good for money, a firm and a union sitting at the negotiation table to sign a labour contract, a couple deciding how to split the intra-household chores, two unfriendly countries trying to reach a lasting peace agreement, or out-of-court negotiations between two litigating parties.

In all these cases three basic ingredients are present: (*a*) the status quo, or the disagreement point, that is, the arrangement that is expected to prevail if an agreement is not reached; (*b*) the presence of mutual gains from cooperation; and (*c*) the multiplicity of possible cooperative arrangements, which split the resulting surplus in different ways.

If the situation involves more than two parties, matters are different, as set out in von Neumann and Morgenstern (1944). Indeed, in addition to the possibilities already identified of either disagreement or agreement among all parties, it is conceivable that an agreement be reached among only some of the parties. In multilateral settings, we are therefore led to distinguish pure bargaining problems, in which partial agreements of this kind are not possible because subcoalitions have no more power than individuals alone, from coalitional bargaining problems (or simply coalitional problems), in which partial agreements become a real issue in formulating threats and predicting outcomes. An example of a pure bargaining problem would be a round of talks among countries in order to reach an international trade treaty in which each country has veto power, whereas an example of a coalitional bargaining problem would be voting in legislatures. In this article we concentrate on pure bargaining problems, leaving the description of coalitional problems to other articles in the dictionary. We are likewise not concerned with the vast informal literature on bargaining, which conducts case studies and tries to teach bargaining skills for the 'real world' (for this purpose, the reader is referred to Raiffa 1982).

Approaches to Bargaining Before Game Theory

Before the adoption of game theoretic techniques, economists deemed bargaining problems (also called bilateral monopolies at the time) indeterminate. This was certainly the position adopted by important economic theorists, including Edgeworth (1881) and Hicks (1932). More specifically, it was believed that the solution to a bargaining problem must satisfy both individual rationality and collective rationality properties: the former means that neither party should end up worse than in the status quo and the latter refers to Pareto efficiency. Typically, the set of individually rational and Pareto-efficient agreements is very large in a bargaining problem, and these theorists were inclined to believe that theoretical arguments could go no further than this in obtaining a prediction. To be able to obtain such a prediction, one would have to rely on extra-economic variables, such as the bargaining power and abilities of either party, their state of mind in negotiations, their religious beliefs, the weather and so on.

A precursor to the game theoretic study of bargaining, at least in its attempt to provide a more determinate prediction, is the analysis of Zeuthen (1930). This Danish economist formulated a principle whereby the solution to a bargaining problem was dictated by the two parties' risk attitudes (given the probability of breakdown of negotiations following the adoption of a tough position at the bargaining table). The reader is referred to Harsanyi (1987) for a version of Zeuthen's principle and its connection with Nash's bargaining theory. The remainder of this article deals with game theoretic approaches to bargaining.

The Axiomatic Theory of Bargaining

Nash (1950, 1953) are seminal papers that constitute the birth of the formal theory of bargaining. Two assumptions are central in Nash's theory. First, bargainers are assumed to be fully rational individuals, and the theory is intended to yield predictions based exclusively on data relevant to them (in particular, the agents are equally skilful in negotiations, and the other extraneous factors mentioned above do not play a role).

Second, a bargaining problem is represented as a pair (S, d) in the utility space, where S is a compact and convex subset of $\mathbb{R}^2$ – the feasible set of utility pairs – and $d \in \mathbb{R}^2$ is the disagreement utility point. Compactness follows from standard assumptions such as closed productions sets and bounded factor endowments, and convexity is obtained if one uses expected utility and lotteries over outcomes are allowed. Also, the set S must include points that dominate the disagreement point, that is, there is a positive surplus to be enjoyed if agreement is reached and the question is how this surplus should be divided. As in most of game theory, by 'utility' we mean von Neumann–Morgenstern expected utility; there may be underlying uncertainty,

perhaps related to the probability of breakdown of negotiations. We shall normalize the disagreement utilities to 0 (this is without loss of generality if one uses expected utility because any positive affine transformation of utility functions represents the same preferences over lotteries). The resulting bargaining problem is called a normalized problem.

With this second assumption, Nash is implying that all information relevant to the solution of the problem must be subsumed in the pair (S, d). In other words, two bargaining situations that may include distinct details ought to be solved in the same way if both reduce to the same pair (S, d) in utility terms. In spite of this, it is sometimes convenient to distinguish between feasible utility pairs (points in S) and feasible outcomes in physical terms (such as the portions of a pie to be created after agreement).

Following the two papers by Nash (1950, 1953), bargaining theory is divided into two branches, the so-called axiomatic and strategic theories. The axiomatic theory, born with Nash (1950), which most authors identify with a normative approach to bargaining, proposes a number of properties that a solution to any bargaining problem should have, and proceeds to identify the solution that agrees with those principles. Meanwhile, the strategic theory, initiated in Nash (1953), is its positive counterpart: the usual approach here is the exact specification of the details of negotiation (timing of moves, information available, commitment devices, outside options and threats) and the identification of the behaviour that would occur in those negotiation protocols. Thus, while the axiomatic theory stresses how bargaining *should* be resolved between rational parties according to some desirable principles, the strategic theory describes how bargaining *could* evolve in a non-cooperative extensive form in the presence of common knowledge of rationality. Interestingly, the two theories connect and complement one another.

The Nash Bargaining Solution

The first contribution to axiomatic bargaining theory was made by John Nash in his path-breaking paper published in 1950. Nash wrote it as a term paper in an international trade course that he was taking as an undergraduate at Carnegie, at the age of 17. At the request of his Carnegie economics professor, Nash mailed his term paper to John von Neumann, who had just published his monumental book with Oskar Morgenstern. John von Neumann may not have paid enough attention to a paper sent by an undergraduate at a different university, and nothing happened with the paper until Nash arrived in Princeton to begin studying for his Ph.D. in mathematics.

According to Nash (1950), a solution to bargaining problems is simply a function that assigns to each normalized utility possibility set S one of its feasible points (recall that the normalization of the disagreement utilities has already been performed). The interpretation is that the solution dictates a specific agreement to each possible bargaining situation. Examples of solutions are: (*a*) the disagreement solution, which assigns to each normalized bargaining problem the point (0,0), a rather pessimistic solution; and (*b*) the dictatorial solution with bargainer 1 as the dictator, which assigns the point in the Pareto frontier of the utility possibility set in which agent 2 receives 0 utility. Surely, neither of these solutions looks very appealing: while the former is not Pareto efficient because it does not exploit the gains from cooperation associated with an agreement, the latter violates the most basic fairness principle by being so asymmetric.

Nash (1950) proceeds by proposing four desirable properties that a solution to bargaining problems should have.

1. *Scale invariance or independence of equivalent utility representations*. Since the bargaining problem is formulated in von Neumann–Morgenstern utilities, if utility functions are re-scaled but they represent the same preferences, the solution should be re-scaled in the same fashion. That is, no fundamental change in the recommended agreement will happen following a re-normalization of utility functions; the solution will simply re-scale utilities accordingly.
2. *Symmetry.* If a bargaining problem is symmetric with respect to the 45 degree line, the

solution must pick a point on it: in a bargaining situation in which each of the threats made by one bargainer can be countered by the other with exactly the same threat, the two should be equally treated by the solution. This axiom is sometimes called 'equal treatment of equals' and it ensures that the solution yields 'fair' outcomes.

3. *Pareto efficiency.* The solution should pick a point of the Pareto frontier. As elsewhere in welfare economics, efficiency is the basic ingredient of a normative approach to bargaining; negotiations should yield an efficient outcome in which all gains from cooperation are exploited.
4. *Independence of irrelevant alternatives* (*IIA*). Suppose a solution picks a point from a given normalized bargaining problem. Consider now a new normalized problem, a subset of the original, but containing the point selected earlier by the solution. Then, the solution must still assign the same point. That is, the solution should be independent of 'irrelevant' alternatives: as in a constrained optimization programme, the deleted alternatives are deemed irrelevant because they were not chosen when they were present, so their absence should not alter the recommended agreement.

With the aid of these four axioms, Nash (1950) proves the following result:

Theorem 1 There is a unique solution to bargaining problems that satisfies properties (1–4): it is the one that assigns to each normalized bargaining problem the point that maximizes the product of utilities of the two bargainers.

Today we refer to this solution as the 'Nash solution'. Although some of the axioms have been the centre of some controversy – especially his fourth, IIA, axiom – the Nash solution has remained as the fundamental piece of this theory, and its use in applications is pervasive.

Some features of the Nash solution ought to be emphasized. First, the theory can be extended to the multilateral case, in which there are $n \geq 3$ parties present in bargaining: in a multilateral problem, it continues to be true that the unique solution that satisfies (1–4) is the one prescribing that agreement in which the product of utilities is maximized. See Lensberg (1988) for an important alternative axiomatization.

Second, the theory is independent of the details of the negotiation-specific protocols, since it is formulated directly in the space of utilities. In particular, it can be applied to problems where the utilities are derived from only one good or issue, as well as those where utility comes from multiple goods or issues.

Third, perhaps surprisingly because risk is not explicitly part of Nash's story, it is worth noting that the Nash solution punishes risk aversion. All other things equal, it will award a lower portion of the surplus to a risk-averse agent. This captures an old intuition in previous literature that risk aversion is detrimental to a bargainer: afraid of the bargaining breakdown, the more risk-averse a person is, the more he will concede in the final agreement. For example, suppose agents are bargaining over how to split a surplus of size 1. Let the utility functions be as follows: $u_1(x_1) = x_1^{\alpha}$ for $0 < \alpha \leq 1$, and $u_2(x_2) = x_2$, where x_1 and x_2 are the non-negative shares of the surplus, which add up to 1. The reader can calculate that the Pareto frontier of the utility possibility set corresponds to the agreements satisfying the equation $u_1^{1/\alpha} + u_2 = 1$. Therefore, the Nash solution awards the utility vector $(u_1^*, u_2^*) = \left(\left(\frac{\alpha}{\alpha+1}\right)^{\alpha}, \frac{1}{\alpha+1}\right)$, corresponding to shares of the surplus $(x_1, x_2) = \left(\frac{\alpha}{\alpha+1}, \frac{1}{1+\alpha}\right)$. Note how the smaller α is, the more risk-averse bargainer 1 is.

Fourth, Zeuthen's principle turns out to be related to the Nash solution (see Harsanyi 1987): in identifying the bargainer who must concede next, the Nash product of utilities of the two proposals plays a role. See Rubinstein et al. (1992) for a related novel interpretation of the Nash solution.

Fifth, the family of asymmetric Nash solutions has also been used in the literature as a way to capture unequal bargaining powers. If the bargaining power of player i is $\beta_i \in [0, 1]$, $\Sigma_i \beta_i = 1$, the asymmetric Nash solution with weights (β_1, β_2) is defined as the function that assigns to each

normalized bargaining problem the point where $u_1^{\beta_1} u_2^{\beta_2}$ is maximized.

The Kalai–Smorodinsky Bargaining Solution

Several researchers have criticized some of Nash's axioms, IIA especially. To see why, think of the following example, which begins with the consideration of a symmetric right-angled triangle S with legs of length 1. Clearly, efficiency and symmetry alone determine that the solution must be the point (1/2, 1/2). Next, chop off the top part of the triangle to get a problem $T \subset S$, in which all points where $u_2 > 1/2$ have been deleted. By IIA, the Nash solution applied to the problem T is still the point (1/2, 1/2).

Kalai and Smorodinsky (1975) propose to retain the first three axioms of Nash's, but drop IIA. Instead, they propose an individual monotonicity axiom. To understand it, let $a_i(S)$ be the highest utility that agent i can achieve in the normalized problem S, and let us call it agent i's aspiration level. Let $a(S) = (a_1(S), a_2(S))$ be the utopia point, typically not feasible.

5. *Individual monotonicity.* If $T \subset S$ are two normalized problems, and $a_j(T) = a_j(S)$, the solution must award i a utility in S at least as high as in T.

We can now state the Kalai–Smorodinsky theorem:

Theorem 2 There is a unique solution to bargaining problems that satisfies properties (1, 2, 3, 5): it is the one that assigns to each normalized bargaining problem the intersection point of the Pareto frontier and the straight line segment connecting 0 and the utopia point.

Note how the Kalai–Smorodinsky solution awards the point (2/3, 1/3) to the problem T of the beginning of this subsection. In general, while the Nash solution pays attention to local arguments (it picks out the point of the smooth Pareto frontier where the utility elasticity $(du_2/u_2)/(du_1/u_1)$ is (1), the Kalai–Smorodinsky solution is mostly driven by 'global' considerations, such as the highest utility each bargainer can obtain in the problem.

Other Solutions

Although the two major axiomatic solutions are Nash's and Kalai–Smorodinsky's, authors have derived a plethora of other solutions also axiomatically (see, for example, Thomson 1994, for an excellent survey). Among them, one should perhaps mention the egalitarian solution, which picks out the point of the Pareto frontier where utilities are equal. This is based on very different principles, much more tied to ethics of a certain kind and less to the principles governing bargaining between two rational individuals. In particular, note how it is not invariant to equivalent utility representations, because of the strong interpersonal comparisons of utilities that it performs.

The Strategic Theory of Bargaining

Now we are interested in specifying the details of negotiations. Thus, while we may lose the generality of the axiomatic approach, our goal is to study reasonable procedures and identify rational behaviour in them. For this and the next section, some major references include Osborne and Rubinstein (1990) and Binmore et al. (1992).

Nash's Demand Game

Nash (1953) introduces the first bargaining model expressed as a non-cooperative game. Nash's demand game, as it is often called, captures in crude form the force of commitment in bargaining. Both bargainers must demand simultaneously a utility level. If the pair of utilities is feasible, it is implemented; otherwise, there is disagreement and both receive 0. This game admits a continuum of Nash equilibrium outcomes, including every point of the Pareto frontier, as well as disagreement. The first message that emerges from Nash's demand game is the indeterminacy of equilibrium outcomes, commonplace in non-cooperative game theory. In the same paper, advancing ideas that would be developed a couple of decades later, Nash proposed a refinement of the Nash equilibrium concept based on the possibility of uncertainty around the true feasible set. The result was a selection of one Nash

equilibrium outcome, which converges to the Nash solution agreement as uncertainty vanishes.

The model just described is referred to as Nash's demand game with fixed threats: following an incompatible pair of demands, the outcome is the fixed disagreement point. Nash (1953) also analysed a variable threats model. In it, the stage of simultaneous demands is preceded by another stage, in which bargainers choose threats. Given a pair of threats chosen in the first stage, the refinement argument is used to obtain the Nash solution of the induced problem in the ensuing subgame (where the threats determine an endogenous disagreement point). Solving the entire game is possible by backward induction, appealing to logic similar to that in von Neumann's minimax theorem; see Abreu and Pearce (2002) for a connection between the variable threats model and repeated games.

The Alternating Offers Bargaining Procedure

The following game elegantly describes a stylized protocol of negotiations over time. It was studied by Stahl (1972) under the assumption of an exogenous deadline (finite horizon game) and by Rubinstein (1982) in the absence of a deadline (infinite horizon game). Players 1 and 2 are bargaining over a surplus of size 1. The bargaining protocol is one of alternating offers. In period 0, player 1 begins by making a proposal, a division of the surplus, say $(x, 1-x)$, where $0 \leq x \leq 1$ represents the part of the surplus that she demands for herself. Player 2 can then either accept or reject this proposal. If he accepts, the proposal is implemented; if he rejects, a period must elapse for them to come back to the negotiation table, and at that time (period 1) the roles are reversed so that player 2 will make a new proposal $(y, 1-y)$, where $0 \leq y \leq 1$ is the fraction of surplus that he offers to player 1. Player 1 must then either accept the new proposal, in which case bargaining ends with $(y, 1-y)$ as the agreement, or reject it, in which case a period must elapse before player 1 makes a new proposal. In period 2, player 1 proposes $(z, 1-z)$, to which player 2 must respond, and so on. The T-period finite horizon game imposes the disagreement outcome, with zero payoffs, after T proposals have been rejected. On the other hand, in the infinite horizon version, there is always a new proposal in the next period after a proposal is rejected.

Both players discount the future at a constant rate. Let $\delta \in [0, 1)$ be the per period discount factor. To simplify, let us assume that utility is linear in shares of the surplus. Therefore, from a share x agreed in period t, a player derives a utility of $\delta^{t-1}x$. Note how utility is increasing in the share of the surplus (monotonicity) and decreasing in the delay with which the agreement takes place (impatience).

A strategy for a player is a complete contingent plan of action to play the game. That is, a strategy specifies a feasible action every time a player is called upon to act in the game. In a dynamic game, Nash equilibrium does little to restrict the set of predictions: for example, it can be shown that in the alternating offers games, any agreement $(x, 1-x)$ in any period t, $0 \leq t \leq T < \infty$, can be supported by a Nash equilibrium; disagreement is also a Nash equilibrium outcome.

The prediction that game theory gives in a dynamic game of complete information is typically based on finding its subgame perfect equilibria. A subgame perfect equilibrium (SPE) in a two-player game is a pair of strategies, one for each player, such that the behaviour specified by them is a best response to each other at every point in time (not only at the beginning of the game). By stipulating that players must choose a best response to each other at every instance that they are supposed to act, SPE rules out incredible threats: that is, at an SPE players have an incentive to carry out the threat implicit in their equilibrium strategy because it is one of the best responses to the behaviour they expect the other player to follow at that point.

In the alternating offers games described above, there is a unique SPE, in both the finite and the infinite horizon versions. The SPE in the finite horizon game is found by backward induction. For example, in the one-period game, the so-called ultimatum game, the unique SPE outcome is the agreement on the split $(1, 0)$: since the outcome of a rejection is disagreement, the responder will surely accept any share of $\in > 0$, which implies that in equilibrium the

proposer ends up taking the entire surplus. Using this intuition, one can show that the outcome of the two-period game is the immediate agreement on the split $(1-\delta, \delta)$: anticipating that if negotiations get to the final period, player 2 (the proposer in that final period) will take the entire surplus, player 1 persuades him not to get there simply by offering him the present discounted value of the entire surplus, that is, δ, while she takes the rest. This logic continues and can be extended to any finite horizon. The sequence of SPE outcomes so obtained as the deadline $T \to \infty$ is shown to converge to the unique SPE of the infinite horizon game. This game, more challenging to solve since one cannot go to its last period to begin inducting backwards, was studied in Rubinstein (1982). We proceed to state its main theorem and discuss the properties of the equilibrium (see Shaked and Sutton 1984, for a simple proof).

Theorem 3 Consider the infinite horizon game of alternating offers, in which both players discount the future at a per period rate of $\delta \in [0, 1)$. There exists a unique SPE of this game: it prescribes immediate agreement on the division $\left(\frac{1}{1+\delta}, \frac{\delta}{\delta+1}\right)$.

The first salient prediction of the equilibrium is that there will not be any delay in reaching an agreement. Complete information – each player knows the other player's preferences – and the simple structure of the game are key factors to explain this.

The equilibrium awards an advantage to the proposer, as expressed by the discount factor: note how the proposer's share exceeds the responder's by a factor of $1/\delta$. Given impatience, having to respond to a proposal puts an agent in a delicate position, since rejecting the offer entails time wasted until the next round of negotiations. This is the source of the proposer's advantage. Of course, this advantage is larger, the larger the impatience of the responder: note how if $\delta = 0$ (extreme impatience), the equilibrium awards all the surplus to the proposer because her offer is virtually an ultimatum; on the other hand, as $\delta \to 1$, the first-mover advantage disappears and the equilibrium tends to an equal split of the surplus.

To understand how the equilibrium works and in particular how the threats employed in it are credible, consider the SPE strategies. Both players use the same strategy, and it is the following: as a proposer, each player always asks for $1/(1 + \delta)$ and offers $\delta/(1 + \delta)$ to the other party; as a responder, a player accepts an offer as long as the share offered to the responder is at least $\delta/(1 + \delta)$. Note how rejecting a share lower than $\delta/(1 + \delta)$ is credible, in that its consequence, according to the equilibrium strategies, is to agree in the next period on a split that awards the rejecting player a share of $1/(1 + \delta)$, whose present discounted value at the time the rejection occurs is exactly $\delta/(1 + \delta)$.

To appreciate the difference from Nash equilibrium, let us argue, for example, that the split (0,1) cannot happen in an SPE. This agreement happens in a Nash equilibrium, supported by strategies that ask player 1 to offer the whole pie to player 2, and player 2 to reject any other offer. However, the threat embodied in player 2's strategy is not credible: when confronted with an offer $(\in, 1-\in)$ for $\delta < 1-\in < 1$, player 2 will have to accept it, contradicting his strategy. Can the reader argue why the Nash equilibrium split (1,0) is not an SPE outcome either (because to do so one would need to employ non-credible threats)? Rubinstein (1982) shows that the same non-credible threats are associated with any division of the pie other than the one identified in the theorem.

The Rubinstein–Stahl alternating offers game provides an elegant model of how negotiations may take place over time, and its applications are numerous, including bargaining problems pertaining to international trade, industrial organization, or political economy. However, unlike Nash's axiomatic theory, its predictions are sensitive to details. This is no doubt one of its strengths because one can calibrate how those details may influence the theory's prediction, but it is also its weakness in terms of lack of robustness in predictive power.

Incomplete Information

In a static framework, Chatterjee and Samuelson (1983) study a double auction. A buyer and a

seller are trying to transact a good. Each proposes a price, and trade takes place at the average of the two prices if and only if the buyer's price exceeds the seller's. Each trader knows his own valuation for the good. However, there is incomplete information on each side concerning the other side's valuation. It can be shown that in any equilibrium of this game there are inefficiencies: given certain *ex post* valuations of buyer and seller, there should be trade, yet it is precluded because of incomplete information, which leads traders to play 'too tough'.

Let us now turn to bargaining over time. As pointed out above, one prediction of the Rubinstein–Stahl model is immediate agreement. This may clash with casual observation; one may simply note the existence of strikes, lockouts and long periods of disagreement in many actual negotiations. As a consequence, researchers have suggested the construction of models in which inefficiencies, in the form of delay in agreement, occur in equilibrium. The main feature of bargaining models with this property is incomplete information. (For delay in agreement that does not rely on incomplete information, see Fernandez and Glazer 1991; Avery and Zemsky 1994; Busch and Wen 1995.)

If parties do not know each other's preferences (impatience rate, per period fixed cost of hiring a lawyer, profitability of the agreement, and so on), the actions taken by the parties in the bargaining game may be intended to elicit some of the information that they do not have, or perhaps to reveal or misrepresent some of the information privately held.

One technical remark is in order. The typical approach is to reduce the uncertainty to a game of imperfect information through the specification of types in the sense of Harsanyi (1967–8). In such games, SPE no longer constitutes an appropriate refinement of Nash equilibrium. The relevant equilibrium notions are perfect Bayesian equilibrium and sequential equilibrium, and in them the off- equilibrium path beliefs play an important role in sustaining outcomes. Moreover, these concepts are often incapable of yielding a determinate prediction in many games, and authors have in these cases resorted to further refinements. One problem of the refinements literature, however, is that it lacks strong foundations. Often the successful use of a given refinement in a game is accompanied by a bizarre prediction when the same concept is used in other games. Therefore, one should interpret these findings as showing the possibilities that equilibrium can offer in these contexts, but the theory here is far from giving a determinate answer.

Rubinstein (1985) studies an alternating offers procedure in which there is one-sided incomplete information (that is, while player 1 has uncertainty regarding player 2's preferences, player 2 is fully informed). Suppose there are two types of player 2: one of them is 'weaker' than player 1, while the other is 'stronger' (in terms of impatience or per period costs). This game admits many equilibria, and they differ as a function of parameter configurations. There are pooling equilibria, in which an offer from player 1 is accepted immediately by both types of player 2. More relevant to the current discussion, there are also separating equilibria, in which player 1's offer is accepted by the weak type of player 2, while the strong type signals his true preferences by rejecting the offer and imposing delay in equilibrium. These equilibria are also used to construct other equilibria with more periods of delay in agreement. Some authors (Gul and Sonnenschein 1988) argue that long delays in equilibrium are the product of strong non-stationary behaviour (that is, a player behaves very differently in and out of equilibrium, as a function of changes in his beliefs). They show that imposing stationary behaviour limits the delay in agreement quite significantly. One advantage of stationary equilibria is their simplicity, but one problem with them is that they impose stationary beliefs (players hold beliefs that are independent of the history of play).

The analysis is simpler and multiplicity of equilibrium is less of a problem in games in which the uninformed party makes all the offers. Consider, for example, a version of the model in Sobel and Takahashi (1983). The two players are a firm and a union. The firm is fully informed, while the union does not know the true profitability of the firm. The union makes all offers in these wage negotiations, and there is discounting across

periods. In equilibrium, different types of the firm accept offers at different points in time: firms whose profitability is not very high can afford to reject the first high wage offers made by the union to signal their private information, while very profitable firms cannot because delay in agreement hurts them too much.

Most papers have studied the case of private values asymmetric information (if a player knows her type, she knows her preferences), although the correlated values case has also been analysed (where knowing one's type is not sufficient to know one's utility function); see Evans (1989) and Vincent (1989). The case of two-sided asymmetric information, in which neither party is fully informed, has been treated, for example, in Watson (1998). In all these results, one is able to find equilibria with significant delay in agreement, implying consequent inefficiencies. Uncertainty may also be about the rationality of the opponent: for example, one may be bargaining with a 'behavioural type' who has an unknown threshold below which he will reject all proposals (see Abreu and Gul 2000).

A more general approach is adopted by studies of mechanism design. The focus is not simply on explaining delay as an equilibrium phenomenon in a given extensive form. Rather, the question is whether inefficiencies are a consequence of equilibrium behaviour in any bilateral bargaining game with incomplete information. The classic contribution to this problem is the paper by Myerson and Satterthwaite (1983). In a bilateral trading problem in which there is two-sided private values asymmetric information and the types of each trader are drawn independently from overlapping intervals, there does not exist any budget-balanced mechanism satisfying incentive compatibility, interim individual rationality and *ex post* efficiency. All these are desirable properties for a trading mechanism. Budget balance implies that payoffs cannot be increased with outside funds. Incentive compatibility requires that each type has no incentive to misrepresent his information. Interim individual rationality means that no type can be worse off trading than not trading. Finally, *ex post* efficiency imposes that trade takes place if and only if positive gains from trade exist. This impossibility result is a landmark of the limitations of bargaining under incomplete information, and has generated an important literature that explores ways to overcome it (see for example Gresik and Satterthwaite 1989; Satterthwaite and Williams 1989).

B

Indivisibilities in the Units

One important way in which Rubinstein's result is not robust happens when there is only a finite set of possible offers to be made (see van Damme et al. 1990; Muthoo 1991). Indivisibilities make it impossible for an exact adjustment of offers to leave the responder indifferent; as a result, multiple and inefficient equilibria appear. The issue concerns how fine the grid of possible instantaneous offers is with respect to the time grid in which bargaining takes place. If the former is finer than the latter, Rubinstein's uniqueness goes through; otherwise it does not. There will be circumstances for which one or the other specification of negotiation rules will be more appropriate.

Multi-Issue Bargaining

The following preliminary observation is worth making: if offers are made in utility space or all issues must be bundled in every offer, Rubinstein's result obtains. Thus, the literature on multi-issue bargaining has looked at procedures that depart from these assumptions.

The first generation of papers with multiple issues assumed that the agenda – that is, the order in which the different issues are brought to the table – was exogenously given. Since each issue is bargained over one at a time, Rubinstein's uniqueness and efficiency result obtains, simply proceeding by backward induction on the issues. Fershtman (1990, 2000) and Busch and Horstmann (1997) study such games, from which one learns the comparative statics of equilibrium when agendas are exogenously fixed. The next group of papers studies more realistic games where the agenda is chosen endogenously by the players. The main lesson from this line of work is that restricting the issues that a proposer can bring to the table is a source of inefficiencies. Inderst (2000) and In and Serrano (2003) study a procedure where agenda is totally unrestricted, that is,

the proposer can make offers on any subset of remaining issues and, by exploiting trade-offs in the marginal rates of substitution between issues, Rubinstein's efficiency result is also found. In contrast, Lang and Rosenthal (2001) and In and Serrano (2004) construct multiple and inefficient equilibria (including those with arbitrarily long delay in agreement) when agenda restrictions are imposed. Finally, Weinberger (2000) considers multi-issue bargaining when the responder can accept selectively subsets of proposals and also finds inefficiencies if issues are indivisible.

Multilateral Bargaining

Even within the case of pure bargaining problems, one needs to make a distinction between different ways to model negotiations. The first extension of the Rubinstein game to this case is due to Shaked, as reported in Osborne and Rubinstein (1990, p. 63); see also Herrero (1985). Today we refer to the Shaked/Herrero game as the 'unanimity game'. In it, one of the players, say player 1, begins by making a public proposal to the others. A proposal is a division of the unit of surplus available when agreement is reached. Players $2, \ldots, n$ then must accept or reject this proposal. If all agree, it is implemented immediately. If at least one of them rejects it, time elapses and in the next period another player, say player 2, will make a new proposal, and so on. Note how these rules reduce to Rubinstein's when there are only two players. However, the prediction emerging from this game is dramatically different. For values of the discount factor that are sufficiently high (if $\delta \geq 1/(n-1)$), every feasible agreement can be supported by an SPE and, in addition, equilibria with an arbitrary number of periods of delay in agreement show up. The intuition for this extreme result is that the unanimity required by the rules in order to implement an agreement facilitates a plethora of equilibrium behaviours. For example, let us see how in the case of $n = 3$ it is possible to sustain an agreement where all the surplus goes to player 3. If player 2 rejects it, the same split will be repeated in the continuation, so it is pointless to reject it. If player 1 changes her proposal to try to obtain a gain, it will be rejected by that responder who in the proposal receives less than 1/2 (there must be at least one). This rejector can be bribed with receiving the entire surplus in the continuation, whose present discounted value is at least 1/2 (recall $\delta \geq 1/2$), thereby rendering his rejection credible. Of course, the choice of player 3 as the one receiving the entire surplus is entirely arbitrary and, therefore, one can see how extreme multiplicity of equilibrium is a phenomenon inherent to the unanimity game. This multiplicity relies on non-stationary strategies, as it can be shown that there is a unique stationary SPE.

An alternative extension of the Rubinstein rules to multilateral settings is given by exit games; see Jun (1987), Chae and Yang (1994), Krishna and Serrano (1996). As an illustration, let us describe the negotiation rules of the Krishna–Serrano game. Player 1 makes a public proposal, a division of the surplus, and the others must respond to it. Those who accept it leave the game with the shares awarded by the proposer, while the rejectors continue to bargain with the proposer over the part of the surplus that has not been committed to any player. A new proposal comes from one of the rejectors, and so on. These rules also reduce to Rubinstein's if $n = 2$, but now the possibility of exiting the game by accepting a proposal has important implications for the predictive power of the theory. Indeed, Rubinstein's uniqueness is restored and the equilibrium found inherits the properties of Rubinstein's, including its immediate agreement and the proposer's advantage (the equilibrium shares are $1/[1+(n-1)\delta]$ for the proposer and $\delta/[1+(n-1)\delta]$ for each responder). Note how, given that the others accept, each responder is *de facto* immersed in a two-player Rubinstein game, so in equilibrium he receives a share that makes him exactly indifferent between accepting and rejecting: this explains the ratio $1/\delta$ between the proposer's and each responder's equilibrium shares. The sensitivity of the result to the exact specification of details is emphasized in other papers. Vannetelbosch (1999) shows that uniqueness obtains in the exit game even with a notion of rationalizability, weaker than SPE; and Huang (2002) establishes that uniqueness is still the result in a model that combines unanimity and exit, since offers can be made both conditional

and unconditional to each responder. Baliga and Serrano (1995, 2001) introduce imperfect information in the unanimity and exit games (offers are not public, but made in personalized envelopes), and multiplicity is found in both, based on multiple off-equilibrium path beliefs. Merlo and Wilson (1995) propose a stochastic specification and also find uniqueness of the equilibrium outcome. In a model often used in political applications, Baron and Ferejohn (1989) study a procedure with random proposers in which the proposals are adopted if approved by simple majority (between the unanimity and exit procedures described).

Bargaining and Markets

Bargaining theory provides a natural approach to understand how prices may emerge in markets as a consequence of the direct interaction of agents. One can characterize the outcomes of models in which the interactions of small groups of agents are formulated as bargaining games, and compare them with market outcomes such as competitive equilibrium allocations. If a connection between the two is found, one is giving an answer to the long-standing question of the origin of competitive equilibrium prices without having to resort to the story of the Walrasian auctioneer. If not, one can learn the importance of the frictions in the model that may be preventing such a connection. Both kinds of results are valuable for economic theory.

Small Markets

Models have been explored in which two agents are bargaining, but at least one of them may have an outside option (see Binmore et al. 1988). Thus, the bargaining pair is part of a larger economic context, which is not explicitly modelled. In the simplest specification, uniqueness and efficiency of the equilibrium is found. In the equilibrium, the outside option is used if it pays better than the Rubinstein equilibrium; otherwise it is ignored. Jehiel and Moldovanu (1995) show that delays may be part of the equilibrium when the agreement between a seller and several buyers is subject to externalities among the buyers: a buyer may have an incentive to reject an offer in the hope of making a different buyer accept the next offer and free- ride from that agreement. In general, these markets involving a small number of agents do not yield competitive allocations because market power is retained by some traders (see Rubinstein and Wolinsky 1990).

Large Markets Under Complete Information

The standard model assumes a continuum of agents who are matched at random, typically in pairs, to perform trade of commodities. If a pair of agents agrees on a trade, they break the match. In simpler models, all traders leave the market after they trade once. In the more general models agents may choose either to leave and consume, or to stay in the market to be matched anew. Some authors have studied steady-state versions, in which the measure of traders leaving the market every period is offset exactly by the same measure of agents entering the market. In contrast, non-steady state models do not keep the measure of active traders constant (one prominent class of non-steady state models is that of one-time entry, in which after the initial period there is no new entry; certain transacting agents exit every period, so the market size dwindles over time). The analysis has been performed with discounting (where δ is the common discount factor that is thought of as being near 1) or without it: in both cases the idea is to describe frictionless or almost frictionless conditions (for example, Muthoo 1993, considers several frictions and the outcomes that result when some, but not all, of them are removed).

The first models were introduced by Diamond and Maskin (1979), Diamond (1981), and Mortensen (1982), and they used the Nash solution to solve each bilateral bargaining encounter. Later each pairwise meeting has been modelled by adopting a procedure from the strategic theory.

The most general results in this area are provided by Gale (1986a, b, c, 1987). First, in a partial equilibrium set-up, a market for an indivisible good is analysed in Gale (1987), under both steady state and non-steady-state assumptions.

The result is that all equilibrium outcomes yield trade at the competitive price when discounting is small: in all equilibria trade tends to take place at only one price, and that price must be the competitive price because it is the one that maximizes each trader's expected surplus. This generalizes a result of Binmore and Herrero (1988) and clarifies an earlier claim made by Rubinstein and Wolinsky (1985). Rubinstein and Wolinsky analysed the market in steady state and claimed that the market outcome was different from the competitive one. Their claim is justified if one measures the sets of traders in terms of the stocks present in the market, but Gale (1987) argues convincingly that, given the steady state imposed on the solution concept, it is the flow of agents into the market every period, not the total stock, that should comprise the relevant demand and supply curves. When this is taken into account, all prices are competitive because the measure of transacting sellers is the same as that of the transacting buyers.

In a more general model, Gale (1986a, b, c) studies an exchange economy with an arbitrary number of divisible goods. Now there is no discounting and agents can trade in as many periods as they wish before they leave the market place. Only after an agent rejects a proposal can he leave the market. Under a number of technical assumptions, Gale shows once again that all the equilibrium outcomes of his game are Walrasian:

Theorem 4 At every market equilibrium, each agent leaves the market with the bundle x_k with probability 1, where the list of such bundles is a Walrasian allocation of the economy.

Different versions of this result are proved in Gale (1986a, c) and in Osborne and Rubinstein (1990). Also, Kunimoto and Serrano (2004) obtain the same result under substantially weaker assumptions on the economy, thereby emphasizing the robustness of the connection between the market equilibria of this decentralized exchange game and the Walrasian allocations of the economy. There are two key steps in this argument: first, one establishes that, since pairs are trading, pairwise efficiency obtains, which under some conditions leads to Pareto efficiency; and second, the equilibrium strategies imply budget balance so that each agent cannot end up with a bundle that is worth more than his initial endowment (given prices supporting the equilibrium allocation, already known to be efficient).

Dagan et al. (2000) also show a Walrasian result, but in their game the trading groups are coalitions of any finite size: in their proof, the force of the core equivalence theorem is exploited. One final comment is pertinent at this point. Some authors (for example, Gale 2000) question the use of coalitions of any finite size in the trading procedure because the 'large' size of some of those groups seems to clash with the 'decentralized' spirit of these mechanisms. On the other hand, one can also argue that for the procedure to allow trade only in pairs, some market authority must be keeping track of this, making sure that coalitions of at least three agents are 'illegal'. Both trading technologies capture appealing aspects of decentralization, depending on the circumstances, and the finding is that either one yields a robust connection with the teachings of general equilibrium theory in frictionless environments. This is one more instance of the celebrated equivalence principle: in models involving a large number of agents, game theoretic predictions tend to converge, under some conditions, to the set of competitive allocations.

Large Markets Under Incomplete Information

If the asymmetric information is of the private values type, the same equivalence result is obtained between equilibria of matching and bargaining models and Walrasian allocations. This message is found, for example, in Rustichini et al. (1994), Gale (1987) and Serrano (2002). In the latter model, for instance, some non-Walrasian outcomes are still found in equilibrium, but they can be explained by features of the trading procedure that one could consider as frictions, such as a finite set of prices and finite sets of traders' types.

The result is quite different when asymmetric information goes beyond private values. For example, Wolinsky (1990) studies a market with pairwise meetings in which there is uncertainty regarding the true state of the world (which determines the true quality of the good being traded).

Some traders know the state, while others do not, and there are uninformed traders among buyers and sellers (two-sided asymmetric information). The analysis is performed in steady state. To learn the true state, uninformed traders sample agents of the opposite side of the market. However, each additional meeting is costly due to discounting. The relevant question is whether information will be transmitted from the informed to the uninformed when discounting is removed. Wolinsky's answer is in the negative: as the discount factor $\delta \to 1$, a non-negligible fraction of uninformed traders transacts at a price that is not *ex post* individually rational. It follows that the equilibrium outcomes do not approximate those given by a fully revealing rational expectations equilibrium (REE). The reason for this result is that, while as $\delta \to 1$ sampling becomes cheaper and therefore each uninformed trader samples more agents, this is true on both sides, so that uninformed traders end up trying to learn from agents that are just as uninformed as they are. Serrano and Yosha (1993) overturn this result when asymmetric information is one-sided: in this case, although the noise force behind Wolinsky's result is not operative because of the absence of uninformed traders on one side, there is a negative force that works against learning, which is that misrepresenting information becomes cheaper for informed traders as $\delta \to 1$. The analysis in Serrano and Yosha's paper shows that, under steady state restrictions, the learning force is more powerful than the misrepresentation one, and convergence to REE is attained. Finally, Blouin and Serrano (2001) perform the analysis without the strong steady-state assumption, and show that with both information structures (onesided and two-sided asymmetries) the result is negative: Wolinsky's noise force in the two-sided case continues to be crucial, while misrepresentation becomes very powerful in the one-sided model because of the lack of fresh uninformed traders. In these models, agents have no access to aggregate market signals; information is heavily restricted because agents observe only their own private history. It would be interesting to analyse other procedures where information may flow more easily.

See Also

- ▶ Nash Program
- ▶ Shapley Value

B

Bibliography

Abreu, D., and F. Gul. 2000. Bargaining and reputation. *Econometrica* 68: 85–118.

Abreu, D., and D. Pearce 2002. *Bargaining, reputation and equilibrium selection in repeated games*. Mimeo, Princeton University.

Avery, C., and P. Zemsky. 1994. Money burning and multiple equilibria in bargaining. *Games and Economic Behavior* 7: 154–168.

Baliga, S., and R. Serrano. 1995. Multilateral bargaining with imperfect information. *Journal of Economic Theory* 67: 578–589.

Baliga, S., and R. Serrano. 2001. Multilateral negotiations with private side-deals: A multiplicity example. *Economics Bulletin* 3 (1): 1–7.

Baron, D., and J. Ferejohn. 1989. Bargaining in legislatures. *American Political Science Review* 83: 1181–1206.

Binmore, K., and M. Herrero. 1988. Matching and bargaining in dynamic models. *Review of Economic Studies* 55: 17–31.

Binmore, K., A. Shaked, and J. Sutton. 1988. An outside option experiment. *Quarterly Journal of Economics* 104: 753–770.

Binmore, K., M. Osborne, and A. Rubinstein. 1992. Noncooperative models of bargaining. In *Handbook of game theory with economic applications*, ed. J. Aumann and S. Hart, vol. 1. New York: Elsevier.

Blouin, M., and R. Serrano. 2001. A decentralized market with common values uncertainty: non-steady states. *Review of Economic Studies* 68: 323–346.

Busch, L.-A., and I. Horstmann. 1997. Bargaining frictions, bargaining procedures and implied costs in multiple-issue bargaining. *Economica* 64: 669–680.

Busch, L.-A., and Q. Wen. 1995. Perfect equilibria in a negotiation model. *Econometrica* 63: 545–565.

Chae, S., and J.-A. Yang. 1994. An *n*-person pure bargaining game. *Journal of Economic Theory* 62: 86–102.

Chatterjee, K., and W. Samuelson. 1983. Bargaining under incomplete information. *Operations Research* 31: 835–851.

Dagan, N., R. Serrano, and O. Volij. 2000. Bargaining, coalitions and competition. *Economic Theory* 15: 279–296.

Diamond, P. 1981. Mobility costs, frictional unemployment, and efficiency. *Journal of Political Economy* 89: 798–812.

Diamond, P., and E. Maskin. 1979. An equilibrium analysis of search and breach of contract I: Steady states. *Bell Journal of Economics* 10: 282–316.

Edgeworth, F. 1881. Mathematical psychics. In *F. Y. Edgeworth's mathematical psychics and further papers on political economy*, ed. P. Newman. Oxford: Oxford University Press. 2003.

Evans, R. 1989. Sequential bargaining with correlated values. *Review of Economic Studies* 56: 499–510.

Fernandez, R., and J. Glazer. 1991. Striking for a bargain between two completely informed agents. *American Economic Review* 81: 240–252.

Fershtman, C. 1990. The importance of the agenda in bargaining. *Games and Economic Behavior* 2: 224–238.

Fershtman, C. 2000. A note on multi-issue two-sided bargaining: Bilateral procedures. *Games and Economic Behavior* 30: 216–227.

Gale, D. 1986a. Bargaining and competition. Part I: Characterization. *Econometrica* 54: 785–806.

Gale, D. 1986b. Bargaining and competition. Part II: Existence. *Econometrica* 54: 807–818.

Gale, D. 1986c. A simple characterization of bargaining equilibrium in a large market without the assumption of dispersed characteristics. Working Paper 86–05. Philadelphia: CARESS, University of Pennsylvania.

Gale, D. 1987. Limit theorems for markets with sequential bargaining. *Journal of Economic Theory* 43: 20–54.

Gale, D. 2000. *Strategic foundations of general equilibrium – Dynamic matching and bargaining games*. Cambridge: Cambridge University Press.

Gresik, T., and M. Satterthwaite. 1989. The rate at which a simple market converges to efficiency as the number of traders increases: An asymptotic result for optimal trading mechanisms. *Journal of Economic Theory* 48: 304–332.

Gul, F., and H. Sonnenschein. 1988. On delay in bargaining with one-sided uncertainty. *Econometrica* 56: 601–611.

Harsanyi, J. 1967–8. Games with incomplete information played by Bayesian players. Parts I, II and III. *Management Science* 14: 159–182, 320–334, 486–502.

Harsanyi, J. 1987. Bargaining. In *The New Palgrave dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 1. London: Macmillan.

Herrero, M. 1985. N-player bargaining and involuntary underemployment. In *A strategic bargaining approach to market institutions*. Ph.D. thesis. London: London School of Economics.

Hicks, J. 1932. *The theory of wages.* London/New York: Macmillan/St. Martin Press, 1963.

Huang, C.-Y. 2002. Multilateral bargaining: Conditional and unconditional offers. *Economic Theory* 20: 401–412.

In, Y., and R. Serrano. 2003. Agenda restrictions in multi-issue bargaining II: Unrestricted agendas. *Economics Letters* 79: 325–331.

In, Y., and R. Serrano. 2004. Agenda restrictions in multi-issue bargaining. *Journal of Economic Behavior and Organization* 53: 385–399.

Inderst, R. 2000. Multi-issue bargaining with endogenous agenda. *Games and Economic Behavior* 30: 64–82.

Jehiel, P., and B. Moldovanu. 1995. Cyclical delay in bargaining with externalities. *Review of Economic Studies* 62: 619–637.

Jun, B. 1987. A structural consideration on 3-person bargaining. In *Essays on topics in economic theory*. Ph.D. thesis. Philadelphia: Department of Economics, University of Pennsylvania.

Kalai, E., and M. Smorodinsky. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43: 513–518.

Krishna, V., and R. Serrano. 1996. Multilateral bargaining. *Review of Economic Studies* 63: 61–80.

Kunimoto, T., and R. Serrano. 2004. Bargaining and competition revisited. *Journal of Economic Theory* 115: 78–88.

Lang, K., and R. Rosenthal. 2001. Bargaining piecemeal or all at once? *Economic Journal* 111: 526–540.

Lensberg, T. 1988. Stability and the Nash solution. *Journal of Economic Theory* 45: 330–341.

Merlo, A., and C. Wilson. 1995. A stochastic model of sequential bargaining with complete information. *Econometrica* 63: 371–399.

Mortensen, D. 1982. Property rights and efficiency in mating, racing, and related games. *American Economic Review* 72: 968–979.

Muthoo, A. 1991. A note on bargaining over a finite set of feasible agreements. *Economic Theory* 1: 290–292.

Muthoo, A. 1993. Sequential bargaining and competition. *Economic Theory* 3: 353–363.

Myerson, R., and M. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.

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

Nash, J. 1953. Two person cooperative games. *Econometrica* 21: 128–140.

Osborne, M., and A. Rubinstein. 1990. *Bargaining and markets.* San Diego: Academic. http://www.economics.utoronto.ca/osborne/bm. Accessed 4 Aug 2005.

Raiffa, H. 1982. *The art and science of negotiation*. Cambridge, MA: Harvard University Press.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

Rubinstein, A. 1985. A bargaining model with incomplete information on time preferences. *Econometrica* 53: 1151–1172.

Rubinstein, A., and A. Wolinsky. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53: 1133–1150.

Rubinstein, A., and A. Wolinsky. 1990. Decentralized trading, strategic behavior and the Walrasian outcome. *Review of Economic Studies* 57: 63–78.

Rubinstein, A., Z. Safra, and W. Thomson. 1992. On the interpretation of the Nash bargaining solution and its extension to non-expected utility preferences. *Econometrica* 60: 1171–1186.

Rustichini, A., M. Satterthwaite, and S. Williams. 1994. Convergence to efficiency in a simple market with incomplete information. *Econometrica* 62: 1041–1063.

Satterthwaite, M., and S. Williams. 1989. The rate of convergence to efficiency in the buyer's bid double auction as the market becomes large. *Review of Economic Studies* 56: 477–498.

Serrano, R. 2002. Decentralized information and the Walrasian outcome: A pairwise meetings market with private values. *Journal of Mathematical Economics* 38: 65–89.
Serrano, R., and O. Yosha. 1993. Information revelation in a market with pairwise meetings: The one-sided information case. *Economic Theory* 3: 481–499.
Shaked, A., and J. Sutton. 1984. Involuntary unemployment as a perfect equilibrium in a bargaining model. *Econometrica* 52: 1351–1364.
Sobel, J., and I. Takahashi. 1983. A multi-stage model of bargaining. *Review of Economic Studies* 50: 411–426.
Stahl, I. 1972. *Bargaining theory*. Stockholm: Stockholm School of Economics.
Thomson, W. 1994. Cooperative models of bargaining. In *Handbook of game theory with economic applications*, ed. R. Aumann and S. Hart, vol. 2. New York: Elsevier.
van Damme, E., R. Selten, and E. Winter. 1990. Alternating bid bargaining with a smallest money unit. *Games and Economic Behavior* 2: 188–201.
Vannetelbosch, V. 1999. Rationalizability and equilibrium in n-person sequential bargaining. *Economic Theory* 14: 353–371.
Vincent, D. 1989. Bargaining with common values. *Journal of Economic Theory* 48: 47–62.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Watson, J. 1998. Alternating-offer bargaining with two-sided incomplete information. *Review of Economic Studies* 65: 573–594.
Weinberger, C. 2000. Selective acceptance and inefficiency in a two-issue complete information bargaining game. *Games and Economic Behavior* 31: 262–293.
Wolinsky, A. 1990. Information revelation in a market with pairwise meetings. *Econometrica* 58: 1–23.
Zeuthen, F. 1930. *Problems of monopoly and economic warfare*. London: Kegan Paul.

Baring Crisis of 1890

Marc Weidenmier

Abstract

The Baring crisis of 1890 is one of the world's most famous financial crises over the last 200 years. The crisis is well known because the Bank of England put together a rescue fund to save the House of Baring. The investment banking firm was in financial trouble because it was the primary debt issuer for Argentina which was experiencing an economic and financial downturn.The Baring crisis is regarded as an early example of a central bank playing the role of a lender of last resort.

Keywords

Argentina; Banking; Baring; Banking crises; Latin America; Lender of last resort

JEL Classifications

N23; G01

The origins of the Baring crisis of 1890, the most famous financial crisis of the 19th century, can be traced to the world debt crisis of 1873 and ensuing recession, which had large economic effects on Argentina and Latin America. The region did not recover from the downturn until the early 1880s following a resurgence of foreign trade and capital flows from Europe. For Argentina, one of the chief obstacles to economic growth and development was the absence of a strong central government. Historically, the national authority shared power with the provincial governments, and it also faced an internal threat from indigenous people who lived on the pampas. The central government consolidated its power and expanded its borders by driving the indigenous people off the pampas in a series of wars during the late 1870s. With the election of Julio Roca as the country's president, the Indian War hero was able to broker an agreement with the ruling elites of the provinces, which centralized the power of the national government.

One of the primary goals of Roca's government was to employ foreign capital to construct railroads, public works, and to modernize Buenos Ayres (Marichal 1989). The new leader's first major loan was a railway issue which completed two major trunk lines in the South American country. The construction of a transportation network throughout the country helped consolidate Roca's power as well as stimulate economic activity by opening up the market for commercial agriculture. Roca also transformed Buenos Aires into the 'Paris of South America' by constructing

broad avenues, spacious parks, a well-functioning water supply and drainage system, and a modern port. The national and local government carried out a series of state-run infrastructure projects in Latin America (Marichal 1989; Mitchener and Weidenmier 2008).

Although the economic policies of the Roca administration stimulated short-run economic activity in Argentina, they posed serious dangers in the long run. The country's expanding debt could only be serviced if the country had sufficient tax revenues. Unfortunately, it would take years before the government would realize significant revenues from commercial activity stimulated by the infrastructure investments (Ford 1956).

Roca finished his term as Argentina's president in 1886. His brother-in-law Miguel Celman became the country's leader following a fraudulent election. Rather than continuing the policies of his predecessor, Celman reduced the government's role in the administration of the railways. The newly elected president sold the Central Norte and Andino railways, two of the country's most important, to British capitalists. The funds from the sale were supposed to be used to reduce the country's rising debt level. Instead of restoring fiscal discipline, the country began issuing additional debt through state banks even though it stopped borrowing funds to finance railway projects.

From 1886 to 1890, Argentina passed a series of 'banking reforms' that fuelled the expansion of credit and paper money issues (Williams 1920). National and provincial banking authorities ratified a Free Banking Law in 1887 which authorized any banking association to issue notes provided it purchased gold bonds to the full amount of the notes issued. There were several problems with the law. It permitted banks meeting minimum capital requirements to issue paper notes backed by government gold bonds. The bank notes, however, were not redeemable in gold, and since the bonds were new issues, they constituted a new liability on the government's balance sheet. The banks that participated in the note issuance scheme floated loans in Europe to finance the purchase of the domestic gold bonds. This scheme worked as long as foreign investors agreed to purchase the Argentine bonds and as long as additional note issuances were backed 100 per cent by specie. By 1890, Argentine provincial banks had issued more than 30 million pounds of external debt.

Argentina's loose monetary and fiscal policies led to a decline in the country's financial and macroeconomic conditions from the mid-1880s until the outbreak of the Baring crisis in 1890. High-powered money grew at an annual average rate of 18 per cent, inflation averaged 17 per cent, and the paper peso depreciated at an average rate of 19 per cent between 1884 and 1890 (della Paolera and Taylor 2001). By 1890, nearly 40 per cent of the foreign borrowing was going towards debt service, and 60 per cent of imports were going toward consumption goods.

Weakening economic conditions in Argentina reduced the demand for Argentine securities on the London market. Domestic investors began to dump the country's paper peso. Although the government used specie to defend the exchange rate, the stock of gold at the Banco Nacional had declined to such an extent by December 1889 that the financial institution could no longer carry out this currency operation. Strikes, demonstrations and a failed coup by military leaders ensued in 1889 and 1890. Inflation reduced the real wages of Argentine workers. The country's lax monetary and fiscal policies drained the banking system of specie, provoked a series of banking crises, and ultimately the Baring crisis in 1890. Argentine real GDP declined by more than ten per cent between 1890- and 1891. Last-minute attempts to reform the country's poor economic policies failed, and the country entered into a decade long recession.

The Argentine crisis potentially had serious implications for global financial markets, especially London. Baring Brothers was the primary investment bank for the South American country. The firm purchased and issued debt for Argentina. The investment bank was heavily involved with the Buenos Aires Water Supply and Drainage Loan, a new debt issue the underwriter failed to sell on the London market (Eichengreen 1999). The House of Baring secretly notified the Bank of England that it could not service its debt

obligations in November 1890. The central bank then pooled financial resources to prevent the beleaguered investment banking firm from causing a larger meltdown on the London market. The Bank of England secured loans from the Bank of France, Russia's central bank, and British financial institutions to help Baring Brothers service its debt obligations and prevent a larger meltdown on the British market. The rescue operation succeeded and prevented a general financial collapse on European markets. Some scholars have argued that the Baring crisis provides one of the earliest examples of a central bank playing the role of a lender-of-last resort in financial markets (Mitchener and Weidenmier 2008).

Although the Bank of England prevented a financial collapse in Europe, the central bank did little to assist Argentina and Latin America. Argentina experienced a deep recession for several years and did not fully recover from the crisis until the turn of the century. In the absence of macroeconomic data such as GDP, Mitchener and Weidenmier (2008) use interest rates to examine the effects of the Baring crisis on emerging markets. They find that interest rates in increased more than 1600 basis points in Latin America while interest rates in other emerging markets were flat. This suggests that the crisis had severe negative macroeconomic effects in Latin America. The evidence suggests that the Baring crisis was largely a regional financial crisis which had few economic effects outside Latin America.

See Also

► Banking Crises

Bibliography

Della Paolera, G., and A. Taylor. 2001. *Straining at the anchor: The Argentine currency board and the search for macroeconomic stability, 1880–1935*. Chicago: University of Chicago Press.

Eichengreen, B. 1999. The baring crisis in a Mexican mirror. *International Political Science Review* 20: 249–270.

Ford, A. 1956. Argentina and the baring crisis of 1890. *Oxford Economic Papers* 8: 127–150.

Marichal, C. 1989. *A century of debt crises in Latin America*. Princeton: Princeton University Press.

Mitchener, K., and M. Weidenmier. 2008. Baring crisis and the Great Latin American meltdown of the 1890s. *Journal of Economic History* 68: 462–500.

Williams, J. 1920. *Argentine international trade under inconvertible paper currency, 1880–1900*. Cambridge, MA: Harvard University Press.

Barone, Enrico (1859–1924)

F. Caffé

Keywords

Barone, E.; Edgeworth, F. Y.; Einaudi, L.; General equilibrium; Indirect taxation; Mathematics and economics; Pantaleoni, M.; Paretian income curve; Pareto, V.; Public finance; Socialism; Tâtonnement; Walras, L.; Wicksell, J. G. K.; Wieser, F. F. von

JEL Classifications

B31

Barone was born in Naples on 22 December 1859 and died in Rome on 14 May 1924. His education provided him with a solid grounding in the classics and in mathematics, with a view to embarking on a military career. He was appointed in 1894 to the Officers' Training School, where he was 'teacher in charge of military history'. He remained in this position until 1902, when he became the head of the historical office of the General Staff, and was given the rank of colonel.

He resigned in 1906, having already published an excellent series of biographical and historical military studies which altered the traditional concept of historical study in that field, by applying to it a method of successive approximation to which his growing interest in economics had introduced him.

His acquaintance with Maffeo Pantaleoni and Vilfredo Pareto provided him with the opportunity

of collaborating with the *Giornale degli Economisti.* This association proved to be extremely valuable and productive and was to last from 1894 right up to the year of his death. It was in this periodical that in September/October 1908 he published the article 'II Ministro della Produzione nello Stato Collettivista'. This article was for a long time considered to be a mere 'curiosum'. However, after its publication in English in a volume edited by Hayek in 1935, it was destined to place its author, together with von Wieser and Pareto, alongside the founders of the pure theory of a socialist economy.

The whole discussion on collective economic planning, as it had developed since the 1920s, had ideological motivations and implications. These were totally excluded from Barone's article. The paper was, above all, a very ingenious illustration of one of Barone's deep beliefs: the usefulness of mathematical tools in clarifying questions which otherwise remain intricate and obscure. In fact it was Barone's use of equations which established the formal equivalence of the basic economic categories between a society based on private ownership in perfectly competitive conditions and a socialist society, in which the distinct need to establish the relative distribution of income was recognized. As Samuelson writes, the innovative meaning of Barone's contribution was that 'by avoiding all mention of utility and indeed without introducing even the notion of indifference curves, Barone was able to break new ground along lines which have in recent years become associated with the economic theory of index numbers'.

The importance of Barone's arguments in the 1930s debate on the economics of socialism in which he used the idea of a Pareto optimum and improved its application, was also not fully appreciated. It remained for Samuelson's *Foundations of Economic Analysis* (1948) to give a complete acknowledgement of Barone's development (adding different products after they have been weighted by their respective prices through a process of tâtonnement) of the Paretian optimum conditions as they relate to the planning of production under collectivism.

In addition to his connections with the economists already mentioned, Barone was acquainted with the famous academics of the time, both Italian and foreign (in particular, Walras) and they all in various ways underlined the enormous potential of Barone's intellect, his clever use of analytical tools, and the extreme clarity of his graphics. Walras, for example, wrote to him saying that

> Providence has singled you out to write the historical review of the various attempts made at mathematical economics over the last centuries, which promise to offer a doctrine which will become generally accepted in the next century. I strongly urge you to recognize this as your vocation and I hope that circumstances will allow you to undertake the task.

Alongside this appreciation, however, is the impression that Barone was overstretching his interests, a feeling which was stated in no uncertain terms by Luigi Einaudi: 'Because of the various vicissitudes of a life torn between activity, journalism, learning and the cinema ... Barone, who was not inclined to laborious and painstaking research, produced far fewer fruits than his supporters had anticipated.' The comment on the cinema refers to the fact that Barone, pressed by financial necessity and using his historical and military background, prepared treatments for the booming early Italian film industry.

This division of interests delayed until 1910 Barone's appointment to a chair in political economy at the Advanced Institute of Economics and Commerce in Rome, which later became the Faculty of Economics and Commerce. But with hindsight it cannot be said that Barone's admirers were justified in 'asking for more'. It is nearer the truth to say that he had not taken the trouble to put together his often very original and therefore extremely important papers on various subjects. As often happens, however, the very fact that his work on the pure theory of socialism received so much international acclaim was the cause for inadequate recognition of his other notable contributions. Of these, the much revised *Principi di Economia Politica* (1908) was an excellent textbook, which, together with the booklet *Moneta e Risparmio* (1920), indicated that dynamic market forces constituted the main area of his intellectual

interest. See also his works entitled *Economia coloniale* (1912), 'I Costi Connessi e l'Economia dei Trasporti' (1921), and 'Sindacati (Cartelli e Trust)' (1921). Of comparable importance are Barone's investigations in the field of financial studies, demonstrating an approach different from that of De Viti de Marco and Einaudi. Barone assumed an autonomous position in as much as he availed himself of Pareto's contributions on the stability of the distribution of incomes, using it as the basis of the distribution of taxes amongst the members of the community. There have been numerous criticisms of the statistical foundation of the Paretian income curve, and even Barone admitted that its shape could undergo change according to variations in social composition. Nevertheless, using its formulation, he provided an inductive basis for the study of a central issue in public finance. Barone's other research of recognized theoretical relevance was on the adverse welfare effects of indirect taxes on taxpayers as compared with direct taxes, for the same given tax returns. Barone was also a severe critic of the alternative versions of the financial theories of savings, in particular that of Edgeworth on minimum saving.

Although Barone was at the centre of the major theoretical debates of his time, he suffered from a conflicting loyalty to the two main formulators of general equilibrium theory, Walras and Pareto. Having been one of the first to grasp the logical aspects of general equilibrium theory, Barone was able to suggest ideas which Walras used to improve his formulation of the production function and the theory of distribution. When Pareto criticized the Walrasian formulation, Barone refrained from taking sides between the two exponents of general equilibrium theory, and as a result Walras refused to recognize the suggestions Barone had given him. Barone himself confided to Wicksell that much of his work had aimed at 'bringing peace' between the two great antagonists. He considered their 'heated disputes' to be 'utterly and completely' deplorable. In spite of this show of fidelity, Barone should not be thought of merely as a follower of Walras and Pareto. As Gustavo del Vecchio, an excellent judge of both Italian and international economic thought, observed,

> Barone understood the deep systematic and critical significance of general equilibrium theory, but because he had been brought up on philosophy and history, he was able to fully appreciate how great were the writers who followed the partial approach, of whom Marshall was pre-eminent. For them, economic science existed only where it could be related to concrete and immediate reality by means of our instruments of observation.

See Also

- ▶ Economic Calculation in Socialist Countries
- ▶ Pareto, Vilfredo (1848–1923)
- ▶ Social Welfare Function

Selected Works

A complete bibliography of Barone's military studies is provided by a symposium on the 50th anniversary of his death, published in the periodical *L'Amministrazione della Difesa,* July/October 1974. The economic works of Enrico Barone have been reprinted in three volumes by Zanichelli (Bologna), 1937. The latest partial reprinting was carried out by Cedam (Padua), 1970. Of his works on economics, see:

1894a. Di alcuni problemi fondamentali per la teoria matematica dell'imposta. *Giornale degli Economisti*, March.

1894b. Sul trattamento di questioni dinamiche. *Giornale degli Economisti,* November.

1895. Studi sull distribuzione. *Giornale degli Economisti*, February/March.

1908a. *Principi di economia politica,* 7th ed. Rome: Sampaolesi, 1929.

1908b. II Ministro della produzione nello stato collectivista. *Giornale degli Economisti*, September/October. Reprinted as 'The ministry of production in the collectivist state.' In *Collectivist economic planning,* ed. F.A. Jaffé. London: Routledge, 1935; translated into many other languages.

1912a. *Economia coloniale.* Rome: Sampaolese.

1912b. Studi de economia e finanza. *Giornale degli Economisti,* April/July.

1920. *Moneta e Risparmio.* Rome: Armani.

1921a. I costi connessi e l'economia dei trasporti. *Giornale degli Economisti,* February.
1921b. Sindacati (cartelli e trust). In *Nuova Collana di Economisti*, vol. 7. Turin: Utet, 1956.

Bibliography

Del Vecchio, G. 1925. L'opera scientifica di Enrico Barone. *Giornale degli Economisti*, November.
Einaudi, L. 1939. *Prefazione di Principi di Economia finanziaria di A. De Viti de Marco*. Turin: Einaudi.
Jaffé, W., ed. 1965. *Correspondence of Léon Walras and related papers*, 3 vols. Amsterdam: North-Holland.
Samuelson, P.A. 1948. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Spinedi, F. 1924. Di un metodo nello studio della scienza sociale. *Rivista di Scienze Sociali e Discipline Ausiliarie*.

Barriers to Entry

Luís M. B. Cabral

Abstract

The precise definition of barriers to entry is controversial; different versions have been proposed over the years. The issue is not one of pure semantics, since evidence of barriers to entry plays an important role in merger review and other areas of antitrust policy. One definition that seems to reflect current thought and practice is as follows: barriers to entry are structural, institutional and behavioural conditions that allow established firms to earn economic profits for a significant length of time.

Keywords

Absolute cost advantage; Antitrust; Asymmetric information; Barriers to entry; Collusion; Contracts as barriers to entry; Excess capacity; Game theory; Limit pricing; Merger analysis; Oligopolistic competition; Patents; Predatory pricing; Product differentiation; Scale economies; Sunk costs

JEL Classification

D4

Scholars usually debate theories, proofs, frameworks and the like. Rarely does controversy arise over a definition, as it does in the case of 'barriers entry'.

Economists tend to agree on the relevant issues, for example, what the market outcome is given a set of assumptions regarding costs, demand, and the nature of competition. So why so much argument over a definition? One answer is that words and definitions play an important role in antitrust analysis. For example, the Federal Trade Commission and U.S. Department of Justice's *Antitrust Guidelines for Collaborations Among Competitors* (2000) suggests that evidence of substantial barriers to entry leads to closer scrutiny of the practice being challenged. Entry conditions play a similar role in other areas of antitrust policy (for example, merger analysis) in the United States, the European Union and other parts of the world. So, like it or not, we must address the issue of what barriers to entry are.

Bain (1956) defined an entry barrier as the set of technology or product conditions that allow incumbent firms to earn economic profits in the long run. Bain identified three sets of conditions: economies of scale, product differentiation, and absolute cost advantages of established firms. Stigler (1968) criticized this approach, especially the idea of scale economies as a barrier to entry. He offered an alternative definition: a production cost that must be borne by an entrant but not by an incumbent.

Both of these approaches are incomplete, as a simple example will show. I will consider a series of different markets with the same structural conditions: a demand $D(p)$ and a technology that consists of a fixed cost F and zero variable costs. In market A, potential entrants sequentially decide whether to pay F, which is sunk; and then active firms compete à la Bertrand. Market B is like market A, but entrants collude at the monopoly price. Market C differs from market A in that potential entrants simultaneously decide whether to pay the fixed cost F,

and moreover F is committed only for a short period of time. Finally, in market D potential entrants first simultaneously commit to their price level for a given short period, and then decide whether to pay the fixed cost F, to which they are committed during the same period as they are committed to price.

All of these scenarios feature the same structural conditions, and so the Bain and Stigler tests would yield the same answer. Under the Bain approach, there would be barriers to entry, namely, the scale economies implied by the fixed-cost technology. Under the Stigler definition, there would be no barriers to entry, since all firms face the same cost conditions. But both approaches would miss the substantial differences between the various markets. In market A, the equilibrium is for the first potential entrant to become a monopolist. In market B, firms will enter to the point where each firm makes zero profits (I am ignoring here the integer constraint). In market C there are multiple Nash equilibria. A reasonable equilibrium is for firms to enter with a probability such that their expected profit is zero. However, with positive probability the outcome of this equilibrium is for one firm to be a monopolist, just as in market A. Finally, in market D the equilibrium is for one firm to enter with a price equal to average cost.

The above example, while simplistic, shows the importance of looking beyond costs and demand to include behavioural conditions. What is the timing of moves – that is, what are firms committed to and for how long? The toughness of oligopolistic competition, one of the key differences across the cases in the above example, is largely the result of the assumed timing of moves. The length of time over which costs are committed (how sunk costs are) is also a crucial factor. In fact, the issue of time reveals an additional limitation of the Bain approach, with its emphasis on the long-run equilibrium. What use is it to know that the long-run equilibrium is a symmetric duopoly if it takes years for an entrant to catch up with an established firm?

If we take these considerations into account, and bear in mind the practical antitrust use of the concept of barriers to entry, a reasonable definition seems to be: *the set of structural, institutional and behavioural conditions that allow incumbent firms to earn economic profits for a significant length of time*. Admittedly, this is a fairly general definition, but necessarily so: the problem with other definitions is that, in attempting to be more specific, they become incomplete and potentially misleading.

Strategic Entry Deterrence

In the analysis of entry conditions and barriers to entry, a greater emphasis was initially placed on structural (or exogenous) entry conditions, such as economies of scale or incumbent cost advantages. The game theory 'revolution' of the 1970s and 1980s, however, shifted the focus to firm behaviour. This led to a coherent story of why structural conditions may turn into barriers to entry. Consider, for example, market A in the above example. If two firms imply zero prices, as the Bertrand assumption and zero variable costs imply, then the equilibrium outcome is for one firm to enter and set a monopoly price, no matter how low F is. However, if price competition is not vigorous (market B), then no matter how high F is incumbent firms never earn economic profits. More generally, it's the combination of entry cost levels, the irreversibility assumption and the oligopolistic competition assumption that, together, lead to a barrier to entry.

Once the game theory apparatus was developed, the number of applications blossomed, frequently with particular models formalizing particular instances of entry barriers endogenously created by incumbents. So in the 1970s DuPont increased its capacity in the titanium dioxide industry as a way to preempt entry or expansion by rival firms. From the 1950s to the 1970s, established firms in the ready-to-eat breakfast cereal industry rapidly increased the number of brands they offered, possibly as an entry pre-emption strategy. In the late 1960s and early 1970s, Xerox developed hundreds of patents that it never used ('sleeping patents'), their purpose being allegedly to make it more difficult for an entrant to challenge its plain-paper photocopy

monopoly. Before the expiry of its patent on aspartame, Monsanto signed exclusive contracts with its major customers of Nutrasweet (Coke and Pepsi), effectively reducing the residual demand to a potential entrant. And so on.

Gilbert (1989) provides an excellent, if slightly dated, survey of the game-theoretic work in this area. What is common to all of these examples of strategic entry deterrence is a prior action by incumbents that decreases the probability of subsequent entry. This may result from an increase in entry costs (Xerox's sleeping patents, Nutrasweet's contracts) or a decrease in the entrant's post-entry profits (Dupont's excess capacity, excess number of cereal brands). In fact, it suffices that the entrant's *beliefs* regarding costs and profits shift in the appropriate direction, even if there is no direct effect. In a world of asymmetric information, a low price by the incumbent may be interpreted as an absolute cost advantage and thus discourage entry; and repeated aggressive reaction to past entry episodes may increase the expectation of aggressive reaction to future entry. So the strategies of limit pricing or predatory pricing may also create barriers to entry.

Conclusion

The game theory revolution had the benefit of revealing the rich interaction between structural conditions and behavioural conditions. But it also complicated the task of deriving a simple, general definition of barriers to entry. In other parts of the field of industrial organization, the reaction to the 'embarrassment of riches' created by game theory has been to focus on particular industries. I believe a similar approach must be taken with respect to the concept of barriers to entry and its application.

See Also

- ▶ Anti-trust Enforcement
- ▶ Contestable Markets
- ▶ Market Structure

Bibliography

Bain, J. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Baumol, W., J. Panzar, and R. Willig. 1982. *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovitch.

Carlton, D. 2004. Why barriers to entry are barriers to understanding. *American Economic Review* 94: 466–470.

Demsetz, H. 1982. Barriers to entry. *American Economic Review* 72: 47–57.

Federal Trade Commission and U.S. Department of Justice. 2000. *Antitrust guidelines for collaborations among competitors*. Online. Available at http://www.ftc.gov/os/2000/04/ftcdojguidelines.pdf. Accessed 18 Jan 2007.

Gilbert, R. 1989. Mobility barriers and the value of incumbency. In *Handbook of industrial organization*, ed. R. Schmalensee and R. Willig. Amsterdam: North-Holland.

McAfee, R.P., H. Mialon, and W. Williams. 2004. What is a barrier to entry? *American Economic Review* 94: 461–465.

Schmalensee, R. 2004. Sunk costs and antitrust barriers to entry. *American Economic Review* 94: 471–475.

Stigler, G.J. 1968. *The organization of industry*. Homewood: Richard D. Irwin.

Sutton, J. 1991. *Sunk costs and market structure*. Cambridge, MA: MIT Press.

von Weizsäcker, C.C. 1980. *Barriers to entry*. New York: Springer.

Barter

Keith Hart

Keywords

Aristotle; Bargaining; Barter; Division of labour; Exchange; Gift-exchange; Money; Plato

JEL Classifications

D0

Barter is a simultaneous exchange of commodities, whether goods or labour services, with bargaining and without using money. It is thus a form of trade in which credit is absent or weak,

where buyers and sellers compete and rates are not fixed, and which lacks an abstract measure of value in exchange or payment.

There is no economy known to ethnographers in which barter is the only means of exchange; but there are some in which it is dominant (for example, Humphrey 1985); and many marginal areas where barter plays a significant role alongside varieties of primitive trade and money transactions. Moreover barter is a major component of international trade, especially between east and west; it is an indispensable business tool of many modern corporations; and, with the rise of computerized exchange in the USA, it has begun to worry the Internal Revenue Service.

None of these contemporary examples, however, captures the interest of economists in barter. For it is as a central plank in the origin myth of classical and neoclassical economics that barter owes its prominence in modern thought. Adam Smith traced the 'wealth of nations' to division of labour:

> This division of labour, from which so many advantages are derived, is not originally the effect of any human wisdom! It is the necessary, though very slow and gradual, consequence of a certain propensity in human nature which has in view no such extensive utility; the propensity to truck, barter and exchange one thing for another. (Smith 1776, I. ii, p. 13)

Linking this propensity to the faculties of reason and speech, Smith draws a line between ourselves and the animals: 'Nobody ever saw a dog make a fair and deliberate exchange of one bone for another with another dog' (1776, I. ii, p. 13). Given such a predisposition, mankind took advantage of differences in geography and skill to establish interdependence through primitive barter. Eventually the difficulties inherent in barter led to the emergence of certain commodities as normal means of exchange and eventually to money proper. Barter, as an expression of a natural human tendency, is thus the forerunner of modern markets based on money. It follows that these markets should be allowed to be self-regulating and spared the interventions of political agents claiming to possess superior 'wisdom'.

The founders of marginalist economics (Menger, Jevons) likewise traced the origins of money to the inefficiency of an earlier stage of barter. Most modern writers on money follow their example. In this they all echo a tradition first established by Plato and Aristotle. The Greek philosophers, however, imagined that, for money to come to express proportionate needs in a complementary division of labour, law rather than nature was required. To sum up the standard economists' myth, a natural propensity to exchange led human beings to establish a division of labour articulated by individualized barter in local markets; eventually long-distance trade evolved and with it more efficient markets based on money. The absence of a guiding political agency is an important feature of this story.

The most elegant refutation of such a construct is made by Polanyi in *The Great Transformation* (1944). He suggests that a more plausible historical sequence is the reverse of the above. Starting from a geographically based division of labour, highly placed political agents trade goods over long distances and routinize means of payment in a process leading to the establishment of money. Local markets are sometimes a spinoff of these channels of *grand commerce,* 'thus eventually, but no means necessarily, offering to some individuals an occasion to indulge in their alleged propensity for bargaining and haggling' (Polanyi 1944, p. 58). Clearly, evolutionary parables should be treated with caution, especially when they fall under one pole or the other of an ideological struggle between liberalism and socialism. Barter is invariably found in an economic context marked by several institutions of exchange. What matters is to identify its structural features in juxtaposition with alternative mechanisms. In the following discussion the evidence for barter in primitive or backward economies will be reviewed, before turning briefly to its revival in capitalist economies. The principal conclusion is that an understanding of barter requires a synthetic approach combining politics and markets.

Grierson's classic article on the silent trade (1903) is a compilation of evidence for barter

without face-to-face contact which captures the early fascination of armchair anthropology with the subject. The first modern fieldwork monograph in anthropology was also devoted to institutions of exchange. In *Argonauts of the Western Pacific* (1922), Malinowski set out to challenge what he took to be prevailing models of 'economic man'. His focus was the *kula,* a system of gift-exchange in the islands near New Guinea, involving armshells and necklaces. Under the cover of such an exchange between local leaders, the common people bartered for goods whose uneven distribution owed much to a geographically based division of labour. In addition maritime and inland villages exchanged fish for vegetables, sometimes through a formal rationing system organized by community leaders, sometimes through individual barter.

Malinowski emphasized the contrast of styles and status honour between ceremonial exchange and ordinary barter, although in the first case cited they were spatially united and in the second were institutional alternatives. The Melanesians were as anxious as the ethnographer to stress their absolute antipathy to confusion of the two extreme forms of exchange. Gift-giving was formal, characterized by generosity and delay of a return (implying credit and trust); barter was informal, characterized by conflict in bargaining and immediacy of return (implying no projection of the relationship into the future). One conferred high social standing, the other low status. In practice, ceremonial exchange is a means of establishing a fragile political order for trade through a transfer of tokens of alliance between leaders whose communities are on a footing akin to war, whereas individual barter and the appearance of hostility intrinsic to price negotiations can only be tolerated in a situation marked by peace and stable social order. Whatever the imputed social psychology, ceremonial exchange is a direct political intervention in the market, barter a manifestation of relatively free commodity exchange. Societies lacking states and money cannot rely exclusively on one form or the other. They must combine gift-exchange and barter pragmatically in response to variable degrees of 'peace for the trade'.

More recently, Humphrey (1985) has linked barter to economic disintegration in the periphery. Her case study of a people living near the Nepal-Tibet border accounts for the dominance of barter by the low supply of money. Being very poor, they cannot afford to keep much wealth in the form of money, preferring to satisfy demand immediately in the one-to-one transactions of barter. Under these circumstances money itself becomes an item of barter. Humphrey relates this temporary phenomenon to a collapse of the local political order which has left the population in a fragmented and individuated state. They have a high level of mutual tolerance but no hierarchy through which to organize inter-local trade as they once did. There is sometimes 'delayed barter' involving more valuable items and the extension of credit between trading partners. But his looks like a weak version of that more formalized trade based on trust which perhaps ought not to be confused with barter. Delay in making a return and associated relations of credit/debt are antithetical to barter; for bargaining is impossible if either party does not have the option of withdrawing from the negotiation.

Recent anthropological research has focused on the tendency of bartered goods to fall into distinct 'spheres of exchange'. In a classic article Bohannan (1955) argues that the Tiv of Nigeria prefer to exchange goods of the same broad category and look down on transactions across the boundaries between such spheres. Subsistence items are distinguished from prestige goods like cattle, slaves, metal bars and cloth. The highest level of exchange involves marriageable women only. In the colonial period money destroyed this compartmentalization of exchange by making conversion between spheres easier. Cultural disruption was the result.

This argument confuses several levels of analysis. First, as Marshall pointed out, utilities are never wholly commensurate: subsistence, luxury and prestige goods cannot be equalized simply by sharing a monetary medium of evaluation. It does not make any sense to ask how many sacks of potatoes an Eton education is worth, even though they both have a money price. Second, there are

clearly problems of conversion in barter between low-bulk, high-value items and high-bulk, low-value items, typically between long-distance trade goods and small agricultural surpluses. Livestock and poultry offer one ready means of conversion, however. Again, nobody likes to sell a hi-fi set in order to pay the groceries bill, but such conversions are known to occur. Third – and most damaging – the main force restricting exchange to separate spheres is political and ideological, not economic in the technical sense. Tiv elders control commerce with the outside world and hold their junior kinsmen on the farm through a monopoly of marriageable women. Colonialism – not money as a fetishized abstraction – undermined that control by introducing markets for the young men's goods and labour.

The absence of money does not in itself present an insurmountable obstacle to efficient exchange. Much the most important precondition for barter lies in the forms of political order (or the lack of it); and it is this which is undermined by modern markets and by the states whose power is essential to their functioning. With this in mind we should consider briefly the survival of barter in the trading institutions of the advanced economies.

Much of the trade between the West and the Communist bloc took the form of barter for the obvious reason that the East could not accumulate hard currency reserves. The end of Communism was also associated with substantial intra-country barter; see barter in transition. Third World countries, such as some West African states, barter the products of an ecological division of labour (meat for grain) owing to a general lack of cash. Such activities are similar to the early trade between political agents emphasized by Polanyi. The multinational corporations have treasuries larger than those of many nations, yet they often choose to barter commodities they would normally be unable to sell in open markets – so many thousand gallons of paint for several months' lease of a Bahamas hotel chain.

The laissez-faire economist's myth of barter as an expression of mankind's innate propensity to exchange ought to be replaced by a more complex historical appraisal of the institution's significance. Barter is an extremely widespread phenomenon, occurring in many times and places as a partial and often temporary solution to the problem of exchange. It is not abolished by money and indeed sometimes transforms money itself into an item of barter; and, if recent trends are a reliable indicator, it may now be undergoing a revival in the West. It was always a mistake to suppose that markets expanded without definite political conditions for their maintenance. Barter too rests on variable political conditions which are as much contemporary as they are primitive.

See Also

- ▶ Barter in Transition
- ▶ Economic Anthropology
- ▶ Exchange

Bibliography

Bohannan, P. 1955. Some principles of exchange and investment among the Tiv. *American Anthropologist* 57: 60–70.

Grierson, P. 1903. The silent trade. In *Research in economic anthropology*, ed. G. Dalton, vol. 3. Greenwich: JAI Press. 1980.

Humphrey, C. 1985. Barter and economic disintegration. *Man* 20 (1): 48–72.

Malinowski, B. 1922. *Argonauts of the Western Pacific*. London: Routledge.

Polanyi, K. 1944. *The great transformation*. Boston: Farrar.

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

Barter and Exchange

F. Y. Edgeworth

Barter, as distinct from exchange, is defined by the absence of money both as a medium of exchange and a measure of value. In the absence of a

measure of value, complicated transactions between several dealers are hardly possible; and accordingly barter is generally characterized by the absence of competition. In the absence of competition bargains are not *determinate* in the same sense as in a perfect market. In the former, unlike the latter, case you might suppose the dispositions of the parties, their demand curves or 'schedules' (Marshall) known, and yet even theoretically be unable to predict what would be the terms of the bargain.

As Jevons says of such a case – with, in the context, unnecessary emphasis on the *indivisibility* of the commodity exchanged.

The equations of exchange will fail … I conceive that such a transaction must be settled upon other than strictly economical grounds. The result of the bargain will greatly depend upon the comparative amount of knowledge of each other's positions and needs which either bargainer may possess or manage to obtain in the course of the transaction (*Theory of Political Economy*, 2nd edn, pp. 130–34).

To which Mr Price adds, 'Nor indeed, did they possess the gift of clairvoyance, would the problem be necessarily solved' (*Industrial Peace*, p. 54). It is important to study this property of barter not so much on account of the rudimentary transactions to which the term is properly confined as for the sake of their analogy to the dealings of monopolists and combinations in advanced societies.

[The subject in question is discussed in the following passages: Auspitz and Lieben, *Theorie des Preises*, p. 381; Edgeworth, *Mathematical Psychics*, pp. 20–56; 'Observations on the Mathematical Theory of Economics', *Giornale degli Economist*, March 1891; Marshall, *Principles of Economics*, 'Note on Barter'; Menger, *Grundsätze*, ch. iv; Price, *Industrial Peace*, pp. 14 and 54; Sidgwick, *Political Economy*, book ii, ch. x. The formation of appropriate conceptions on the subject is aided by those economists who, improving on the ordinary 'Robinsonnade', introduce a *second* primitive economic man. Good examples occur in Courcelle Seneuil's *Traité théorique et pratique*, and Mr. Gonner's textbook of *Political Economy*].

References

Auspitz, R., and R. Lieben. 1887. *Zur Theorie des Preises*. Leipzig: Dunker & Humblot.

Courcelle-Seneuil, J.G. 1858. *Traité théorique et pratique d'économie politique*. Paris: Amyot.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Edgeworth, F.Y. 1891. Observations on the mathematical theory of economics. *Giornale degli Economisti*.

Gonner, E.C.K. 1888. *Political economy.* London: R. Sutton & Co.

Marshall, A. 1890. Note on barter. In his *Principles of economics.* London: Macmillan.

Menger, C. 1871. *Grundsätze der Volkwirtschaftslehre*. Vienna: Wilhelm Braumuller.

Sidgwick, H. 1883. *Principles of political economy.* London: Macmillan.

Barter in Transition

Barry W. Ickes

Abstract

One of the striking features of the transition in Russia was the enormous growth in the use of barter and other non-monetary means of payment. The transition from command initially led to a monetization of the economy, but a subsequent re-demonetization was a surprise. Barter was a passing phase in most transition economies but became endemic in Russia. Barter proliferated as inflation was tamed and reached its zenith prior to the August 1998 financial crisis. Various theories of why barter exploded in Russia are discussed and empirical findings are assessed.

Keywords

Arrears; Barter; Barter in transition; Contract enforcement; Credit rationing; Double coincidence of wants; Hold-up problem; Inflation; Institutional trap; Liquidity constraints; Market power; Multilateral barter; Natural monopoly; Price discrimination; Relational capital; Tax evasion; Tax offsets; Trust; Virtual economy

JEL Classifications
P

One of the striking features of the transition in Russia was the enormous growth in the use of barter and other non-monetary means of payment. In addition to conventional barter – goods exchanged for goods – non-monetary transactions were also prevalent in this period. These involved non-monetary IOUs, *veksels,* which were claims on goods from other enterprises or offsets on future taxes. (The literature often treats these as equivalent, and indeed, they do arise from similar causes, but the nature of the transactions is clearly distinct.) What was a passing phase of transition in Central Europe became, by 1997, an endemic feature of the Russian situation. The explosion in barter culminated in the August 1998 Russia crisis, and since then the importance of barter has declined.

The growth in the use of barter has been characterized as 're-demonetization' (Ickes et al. 1997). The Soviet economy (with the partial exception of the household sector) was essentially a non-monetary economy. Central planners' decisions, not purchasing power, determined the production and allocation of goods and services. Money was mainly a record-keeping instrument. A main objective of economic reform was to transform the economy from a partially demonetized planned economy to a monetized market economy. Hence the growth in barter represented a return to a non-monetary economy, or a re-demonetization. By 1997 barter accounted for nearly half of all enterprise transactions: see Aukutsionek (1998), Commander and Mummsen (2000) and Noguera and Linz (2006). Not only was barter used in payments between enterprises (estimates of the share of barter in inter-enterprise transactions ranged from 30 per cent to 80 per cent) but it was also widely used in paying taxes to local, regional, and even federal governments. Even wages were occasionally paid in kind.

The emergence of barter in Russia in the mid-1990s presents a challenge to economic theory. Textbook analysis suggests that barter is inferior to monetary exchange. Barter requires a double coincidence of wants and hence is more costly than monetary transactions. Moreover, in Russia barter exploded as inflation was declining. Hence, the growth of barter was not the result of a flight from money as its store-of-value services declined. Indeed, one indication of this is the fact that this explosion in barter was almost exclusively within the enterprise and budget sectors of the economy. Households were typically involved only to the extent they received wages in kind. This suggests that the growth of barter had something to do with what was happening to enterprises.

Explanations of the prevalence of barter in Russia and other transition economies tend to divide into two types. One group of explanations focuses on circumstances external to the firm and views barter as an involuntary decision. The other group of explanations views the use of barter as a strategic decision by the enterprise to reduce its costs or increase its profitability (survivability).

Barter as a Passive Response

A leading argument of the passive theory views barter as the result of a lack of liquidity. Enterprises engage in barter because they simply lack the cash to use money. This could be due to underdeveloped financial systems (Hendley et al. 1998) or to the effects of macroeconomic tightening (Commander and Mummsen 2000; Noguera and Linz 2006). In either case, the premise is that barter will only be used by enterprises that cannot afford to pay with cash; that is, barter is the result of a liquidity constraint. Hence, as argued in Woodruff (1999), barter is an instrument for cutting prices to enterprises that cannot pay the nominal price for inputs using money. Barter thus allows production to continue for those enterprises that are liquidity constrained. Barter is thus an instrument used to price discriminate. Models with this feature are developed by Ericson and Ickes (2001) and Guriev and Kvasov (2004). The liquidity explanation of barter has the advantage of getting the timing right: barter began to increase as real interest rates rose in response to the switch in policy from monetization to borrowing to finance fiscal deficits. Most of the empirical support for the theory, however, comes from

survey responses of directors who state that they accept barter because their customers lack liquidity. That is, the information on the buyers' lack of liquidity stems from surveys of sellers. A problem with this evidence, however, is that, if it is advantageous to the buyer to pay with goods rather than with money, then buyers will act strategically. That is, they will pretend to be liquidity constrained when they may not be, in order to qualify for barter. What sellers observe is the financial condition that the buyers want the sellers to believe. Hence, the liquidity of an enterprise may be endogenous. If enterprises act strategically, then the seller's information may not be the most accurate indicator of the liquidity position of the buyer.

Some empirical evidence that casts doubt on the liquidity hypothesis comes from a study by Guriev and Ickes (2000) that avoids the problem of uninformed sellers and strategic buyers. To get around the problem of strategic signalling, Guriev and Ickes matched data on the proportions of revenues in cash and non-cash form taken from a survey of directors with the *Goskomstat* database of Russian enterprises, which contains the financial accounts of all large and medium-size industrial enterprises in Russia. This allowed them to compare the share of non-cash payments with the enterprise's financial position. They could find no discernible relationship between the use of barter and the financial condition of the enterprise. The only explanatory variable they found that predicted barter was share of export sales. (This also, perhaps, explains why barter fell dramatically when the ruble depreciated and exports increased.) Most interestingly, they found that the best predictor of whether an enterprise would use barter was lagged barter. This suggests that barter was an institutional trap. Once non-cash payments became a widespread phenomenon, it became part of the strategies of all agents. As barter proliferated it became a 'normal' way of doing business.

Barter as a Choice

The notion that barter is a choice that an enterprise makes presumes that it results in a lowering of its net costs of production or an increase in its net revenues. Employing barter clearly increases the costs of transactions, so it must have some other offsetting benefits. For example, it may afford the buyer the opportunity to pay an effectively lower price, or it may enable enterprises to avoid taxation or reduce the cost of paying taxes. This still begs the question of why the seller is willing to accept lower-priced goods. Presumably, the key reason is the ability to pass these off for payment in taxes. This begs the further question of why governments are willing to allow tax offsets. The prevalence of tax offsets, especially at the regional level, is an accepted fact. But the motivation is more complex. (See Gaddy and Ickes 2002 for a discussion.) Barter may also be used as a means of hiding revenues and avoiding restructuring (see virtual economy).

If we suppose that the effective price of purchasing inputs is cheaper using barter, it follows that enterprises will prefer to pay with barter than with money. There must be some way for sellers to limit the use of barter. One method would be to limit barter to enterprises with which there are good relations (see virtual economy for a discussion of relational capital and its importance in the Russian economy). Indeed, as it may be more difficult to enforce contracts using barter, a high level of relational capital or trust may be needed to enable barter to occur. An alternative method is price discrimination by those with market power.

Barter and Tax Evasion

One reason why enterprises may prefer to use barter is that it reduces the effective burden of taxation. In Russia, the traditional banking system served as a key part of the tax collection system. An enterprise in tax arrears would have its bank account blocked, and all receipts would go directly to the tax service. Such an enterprise thus faced 100 per cent marginal tax rates on revenues paid with money. Monetary transactions between enterprises in Russia were required by law to operate through the banking system. Cash withdrawals could only be made for payment of wages and other incidental uses. Using barter allowed a seller in tax arrears to receive payment

and circumvent the tax authorities. Hence, for such enterprises sufficient surplus would be generated by barter to offset the costs.

Evidence on the role of tax evasion as a motivation for barter is mixed. Some studies (for example, Hendley et al. 1998) find survey evidence in favour of the tax-evasion hypothesis, while others do not (for example, Commander and Mummsen 2000). But in most cases these studies focus the question too narrowly. They typically ask whether enterprises use barter to evade taxes. A more appropriate question would ask whether enterprises use barter to reduce the effective tax burden. Enterprises often use barter not to evade taxes but in order to pay taxes, only in a way advantageous to the enterprise. This is the practice of tax offsets.

The practice of using tax offsets as a means of reducing tax incidence became widespread prior to the 1998 crisis and was a key feature of the virtual economy (see virtual economy). Consider, for example, an enterprise that is able to supply the local government with services in lieu of taxes. The enterprise could pay its tax liability in money, but this would require selling its output for cash. Alternatively, the enterprise can negotiate with the government to supply some service as an offset for taxes. If the enterprise has resources that are not fully utilized, the latter alternative is likely to reduce the effective tax burden on the enterprise. Gaddy and Ickes (2002) provide an abundance of examples of the use of tax offsets.

Any comprehensive theory of barter in Russia in the 1990s must also explain one particularly vexing question: why governments are willing to accept tax payments in kind. It is easy to understand why enterprises would want to pay taxes in kind: this lowers the burden of their payments. It is harder to understand why governments would be willing to accept in-kind payments of taxes.

One explanation for the government's willingness to participate in barter is the virtual economy thesis. The proliferation of tax offsets is a mechanism for the distribution of subsidies in a non-transparent manner. Although more costly than a cash distribution of subsidies, non-transparency provides a more durable means of providing subsidies. They are less likely to be attacked as wasteful. This is especially true when subsidies are distributed through production, by keeping open enterprises that ought to be shut down. Thus it may be in the interest of government officials to keep subsidies non-transparent (see virtual economy).

B

Multilateral Barter

A key problem with barter is the difficulty in finding a double coincidence of wants. It is thus interesting that in Russia multilateral barter chains appeared. Barter was often not bilateral, but part of a chain (see Humphrey 2000). As one report described it:

> The barter chain itself turned out to be a special kind of consumer of the output. But its needs differed from the needs of liquid demand. The barter chains frequently reminded one of the 'production for production's sake' of the [Soviet] planned economy, when a quasi-cooperation gave rise to closed autonomous systems that served only themselves. In a number of enterprises which we surveyed, the share of output necessary simply to support the viability of the chain itself was as high as 30 per cent. (Institute of the Economy in Transition, cited in Gaddy and Ickes 2002)

Thus enterprises engaged in production of goods that were useful for maintaining the barter chain. The network character of barter also means that a web of relationships is crucial to maintaining it. This implies that barter was a conservative force, preserving relationships among enterprises.

Barter and Market Power

A robust finding among students of barter in Russia is that the large natural monopolies (Gazprom, UES, and the State Railways system, *tri tolstayaka,* 'The Three Fat Boys') were heavily involved with barter. This suggests that price discrimination may be a motive for barter. Guriev and Kvasov (2004) develop a model where firms can choose to pay in cash or in barter, and natural monopolies use barter to engage in price discrimination across customers. Unlike the model of Ericson and Ickes (2001) the Guriev–Kvasov model does not require the natural monopolies to receive any benefit from the government in exchange for the lower prices it charges to

low-profitability purchasers. Rather, barter simply facilitates price discrimination and is thus profitable for monopolists. Barter allows enterprises with market power to extract higher prices from those that can afford to pay more. Of course, such discrimination can only occur if markets are not competitive.

Guriev and Ickes (2000) tested the predictions of this model and found that the use of barter increases with concentration. Industries where market concentration is very low display lower prevalence of barter than in other industries. Similarly, larger enterprises that operate in concentrated industries (and do not sell to foreign markets) are much more likely to engage in barter. Similar findings with respect to Russia (but not to Central Europe) were found in an EBRD study (Carlin et al., pp. 247–8).

Barter and Efficiency

As barter is costly it is often assumed that the welfare effects of widespread barter are negative. Barter is typically viewed as a means of avoiding restructuring. An enterprise that successfully restructures may be unable to credibly signal that it is in distress, and thus it may be forced to use cash instead of barter. Ericson and Ickes (2001) developed a general equilibrium model where a restructuring trap exists: enterprises refuse to restructure because they are afraid of losing the benefits of cheap energy supplied via barter. Indeed, a form of this mechanism is at work in most price discrimination models of barter (for example, Guriev and Kvasov). Guriev and Ickes (2000) found empirical support for this hypothesis: in their sample an increase in the share of barter resulted in a decrease in labour productivity.

If barter is the result of liquidity problems external to the enterprise then access to this technology can be welfare enhancing (Noguera and Linz 2006). The basic idea is that in a credit-rationing equilibrium higher interest rates do not provide access to capital; so cash-poor firms that have no access to barter may have to reduce production when real interest rates rise due to crowding out. With access to barter, however, they can maintain production. Of course, to evaluate the welfare consequences one must examine why the enterprises are cash poor in the first place. If this is purely external to the firm then higher production is welfare improving. If the reason they are cash poor is that they produce goods that destroy value then barter actually is welfare decreasing (see virtual economy).

It has also been argued that barter enhances efficiency in an environment of weak contract enforcement. Marin et al. (2000) argue that barter creates 'deal-specific collateral'. They argue that this alleviates the hold-up problem that appears when credit enforcement is prohibitively costly. In such environments transactions that are mutually beneficial take place via barter but would not take place if cash were required. They argue that barter 'is a *self-enforcing arrangement* which makes intermediate producers along the chain of production lose from reneging on the contract' (2000, p. 222). The main difficulty with this theory, however, is to understand how barter creates deal-specific collateral. Presumably, an enterprise can always pledge collateral, and a promise to trade the good to a supplier to is no more credible than a promise to deliver the good if a loan cannot be repaid. The key point is that relational capital among enterprises supports barter, but barter itself does not create relational capital (see Gaddy and Ickes 2002). The agreement between a buyer and a seller to engage in barter does not preclude the buyer from defecting anymore than a pledge of collateral to a supplier would. Thus, it is not easy to see how barter enhances transactions possibilities (though one can see how this might work with *veksels:* see below).

Veksels

As barter is costly, Russian enterprises developed an alternative institution, the use of non-monetary IOUs, or *veksels.* These were claims on output or offsets of future taxes, and their use proliferated prior to the August 1998 crisis. These promissory notes, issued by commercial banks, governments and enterprises, serve as an alternative medium of exchange. The use of *veksels* has become widespread: by one estimate the outstanding stock of these instruments had grown by the spring of 1997

to be roughly two-thirds of the value of all rubles in circulation (ruble M2) (OECD 1997, p. 178). Enterprise *veksels* are issued by large established firms (for example, Gazprom, UES). These notes circulate among chains of enterprises that owe goods to the issuer. Eventually the note is redeemed by some customer of the issuer.

Veksels had two important characteristics that were similar to conventional barter. First, by operating out of the normal channels of the banking system they enabled enterprises to avoid taxation. Second, the use of *veksels* had the effect of keeping enterprises as part of a chain of production. The value of a *veksel* would be much lower outside the chain; hence, they had the effect of keeping enterprises from defecting. A *veksel,* for example, would be issued by a bank to support transactions among suppliers in a chain of production. If one of the suppliers chose not to produce the inputs but defect with the credit the discount on the paper may be quite large. If the credit had been issued in cash, on the other hand, it would be much easier to defect from the production chain. Hence, *veksels* may have served as a means of preserving production relations and extending credit with weak contract enforcement possibilities (Hendley et al. 1998).

Consequences of Barter

Barter raises the private costs of transactions for those engaged in it. Barter becomes prevalent when the institutional and macroeconomic environment is such that it is profitable for enterprises to bear these costs. Hence, it is not barter per se, but the institutional and environmental constraints that generate it that are the problem. The fact that barter locks enterprises into a chain of production and inhibits restructuring is costly to the economy. But it is not the barter that is the cause of the problem, but rather a result of the peculiar economic conditions that make such an equilibrium sustainable.

After the Russian crisis, as the ruble depreciated in real terms and oil prices recovered, the barter equilibrium seems to have broken down. Cash transactions became less costly than they were prior to the crisis. Enhanced government revenues, due to tax reforms and export revenues, led to a decline in tax offsets. Hence, the relative cost of barter increased. The economy re-monetized. Whether barter will return if economic conditions return to their mid-1990s setting is an open question.

See Also

- ▶ Arrears
- ▶ Institutional Trap
- ▶ Soft budget Constraint
- ▶ Virtual Economy

Bibliography

Aukutsionek, S. 1998. Barter in Russian industry. *Voprosy Economiki* 70: 51–60.

Carlin, W., S. Fries, M. Schaffer, and P. Seabright. 2000. Barter and non-monetary transactions in transition economies: Evidence from a cross-country survey. In *The vanishing rouble*, ed. P. Seabright. New York: Cambridge University Press.

Commander, S., and C. Mummsen. 2000. The growth of non-monetary transactions in Russia: Causes and effects. In *The vanishing rouble*, ed. P. Seabright. New York: Cambridge University Press.

Ericson, R.E., and B.W. Ickes. 2001. A model of Russia's virtual economy. *Review of Economic Design* 6: 185–214.

Gaddy, C., and B.W. Ickes. 2002. *Russia's virtual economy*. Washington, DC: Brookings Institution Press.

Guriev, S., and B.W. Ickes. 2000. Barter in Russian firms. In *The vanishing rouble*, ed. P. Seabright. New York: Cambridge University Press.

Guriev, S., and D. Kvasov. 2004. Barter for price discrimination. *International Journal of Industrial Organization* 22: 329–350.

Hendley, K., B.W. Ickes, and R. Ryterman. 1998. Remonetizing the Russian economy. In *Russian enterprise reform: Policies to further the transition*, ed. H.G. Broadman. Washington, DC: World Bank.

Humphrey, C. 2000. An anthropological view of Barter in Russia. In *The vanishing rouble*, ed. P. Seabright. New York: Cambridge University Press.

Ickes, B.W., P. Murrell, and R. Ryterman. 1997. End of the tunnel? The effects of financial stabilization in Russia. *Post-Soviet Affairs (formerly Soviet Economy)* 13: 105–133.

Marin, D., D. Kaufmann, and B. Gorochowskiy. 2000. Barter in transition economies: Competing explanations confront Ukrainian data. In *The vanishing rouble*, ed. P. Seabright. New York: Cambridge University Press.

Noguera, J., and S.J. Linz. 2006. Barter, credit and welfare: A theoretical inquiry into the barter phenomenon in Russia. *Economics of Transition* 14: 719–745.

OECD (Organisation for Economic Co-operation and Development). 1997. *Economic surveys, Russian Federation*. Paris: OECD.
Seabright, P. 2000. *The vanishing rouble*. New York: Cambridge University Press.
Woodruff, D. 1999. *Money unmade: Barter and the fate of Russian capitalism*. Ithaca/London: Cornell University Press.

Barton, John (1789–1852)

Maxine Berg

Keywords

Barton, J.; Capital accumulation; Circulating capital; Colonization; Corn Laws; Fixed capital; Huskisson, W.; Labour supply; London Statistical Society; Machinery question; Malthus, T. R.; Malthus's theory of population; McCullough, J. R.; Monopoly; Old Poor Law; Population; Ricardo, D.; Sismondi, J. C. L. S. de; Smith, A.; Technical change

JEL Classifications

B31

Barton is remembered in the history of economic thought for an early critical discussion of the impact of machinery on employment. A Sussex landowner, he combined an interest in statistical observation with a special concern for the impact of industrial and agrarian change on the condition of the labourer. He was the author of two important books, *Observations on the Circumstances which Influence the Condition of the Labouring Classes of Society* (1817) and *An Inquiry into the Causes of the Progressive Depreciation of Agricultural Labour in Modern Times* (1820). Later, in the 1830s, he wrote several tracts on the Corn Laws and on population and colonization. He was elected a fellow of the London Statistical Society in 1847 and read a paper in 1849, 'The Influence of the Subdivision of the Soil on the Moral and Physical Well-being of the People of England and Wales'. His early manuscript essays show a wide and careful grounding in political economy based on Hume, Smith and Ricardo. His first books were, however, written as interventions in the contemporary debates on the Poor Laws.

Barton's primary purpose in writing both the *Observations* and the *Inquiry* was to challenge Malthusian population theory, and the prevailing opinion that the cause of excess population and falling wages was the support offered by the Old Poor Law. Barton combined abstract reasoning with statistical data in a critique of Malthus and Ricardo that so impressed Schumpeter that he judged it 'a remarkable performance … far above the rest of the literature that currently criticized the class leaders for their lack of realism, actual or supposed'.

Barton drew on population figures from the 16th to the 18th century to challenge Malthusian propositions of the dependence of population growth on levels of capital accumulation. Using data gathered from the agricultural districts, he also challenged assumptions of flexible supplies of labour in response to wage changes. His data provided no support for those who feared that population growth would follow on high wages. Custom and employment prospects, not changing wage rates, were the most important determinant of the age of marriage. Barton dissected the gap between population and labour supply, analysing age structure, apprenticeship, skills and labour immobility. His demographic work impressed Sismondi and induced McCulloch to give up Malthusianism.

The most influential analysis of the *Observations,* however, was Barton's critique of Ricardo's and Malthus' early optimistic assumptions of the impact of capital accumulation and machinery on the working classes. Another reason why high wages could not be blamed for inducing population growth, he argued, was that capital accumulation did not necessarily entail increases in employment. Capital had to be disaggregated into fixed (technological) and circulating (wage goods) capital before its impact on the labour market could be assessed. The demand for labour was dependent on circulating, not fixed, capital. And if wage rates rose relative to commodity prices, employers would substitute machinery for labour. The process

of capital accumulation could, therefore, entail the release of rather than the demand for labour, and the amount of labour employed in the construction and repair of new machinery would provide only small compensation.

Barton's *Observations* was read by political economists and policy-makers – Huskisson and Malthus noted it, Sismondi praised it and McCulloch reviewed it. It was said to have induced Ricardo to make an about-turn in the third edition of his *Principles* and so to write his controversial chapter on machinery accepting the idea that the introduction of machinery could hurt the interests of manual labour. But Ricardo did not introduce this change until the third edition in 1821, and his analysis was rather different. Accepting Barton's point that the introduction of machinery might be induced by wage increases, he added his own novel analysis of autonomous technical change. It is likely that Ricardo changed his views on machinery not because he read Barton but because of contemporary political concern over the machinery issue combined with a timely reminder of Barton's work in a recent correspondence he had with McCulloch.

Barton's later pamphlets and newspaper articles of the 1830s and 1840s extended his early analysis into a general critique of industrialism. He defended the Corn Laws, arguing that labour thrown out of agriculture could not be transferred easily to manufacturing, and that the extension of manufacturing and machinery only concentrated wealth in fewer hands. He drew attention to an Adam Smith forgotten by his contemporaries – the Smith who conducted a radical critique of the monopoly spirit of merchants and manufacturers. John Barton's critique of industrialism and the introduction of machinery was a striking example of a special early 19th-century combination of traditional landed opinion with a radical concern for the condition of labour.

Selected Works

1817. *Observations on the circumstances which influence the condition of the labouring classes of society.* London.

1820. *An inquiry into the causes of the progressive depreciation of agricultural labour in modern times.* London.

1962. *John Barton (1789–1852), economic writings,* 2 vols, ed. G. Sotiroff. Regina: Lynn Publishing Company.

Basics and Non-basics

Neri Salvadori

The distinction between basic and non-basic commodities was introduced by Sraffa (1960). He first provided a definition valid for the single production case and then provided a general definition for the joint production case. In this entry single-output production is assumed except for a few remarks on the distinction between basics and non-basics in the joint production case.

Sraffa calls *basic* a commodity which enters directly or indirectly into the production of all commodities. He calls *non-basic* a commodity which does not have this property. The distinction is important since it is possible to prove that:

(a) basic commodities are indispensable; that is, they need to be produced whatever net output is produced (non-basic commodities are not indispensable and in particular are not produced if the net output consists only of basics);
(b) if the price of a non-basic is changed (i) because of a specific tax on it or (ii) because its method of production is changed, then not all prices are affected, and, specifically, those of basics are not (if the price of a basic is changed, then any other price is changed);
(c) the Standard ratio is defined by the technology of basics only;
(d) relative prices of basics are defined and positive for all non-negative rates of profit lower that the Standard ratio (some non-basics may not have this property);

(e) if the *numéraire* includes only basic commodities, then the relationship between the wage rate and the rate of profit is determined by the technology of basics only;
(f) all basics and none of the non-basics enter into the Standard commodity.

All these statements can easily be shown after that some preliminaries are introduced.

Let us assume that there exist n commodities and n processes to produce them; each process produces *one* commodity. Let a_{ji} be the amount of commodity j used to produce one unit of commodity i; and let l_i be the amount of labour utilized to produce one unit of commodity i. Let $A = [aij]$ and $l = (l_1, l_2, \ldots, \ln)^T$.

Commodity j *enters directly* into the production of commodity i if and only if

$$a_{ij} > 0;$$

commodity j *enterwzm s directly or indirectly* into the production of commodity i if and only if there is a sequence $i_1, \ldots, i_z$ of indices such that

$$a_{ii1}a_{i1i2}\cdots a_{i_zj} > 0$$

that is, if and only if there is a natural number z such that $e^{\mathrm{T}}{}_iA^ze_j > 0$, where e_i and e_j are the ith and the jth unit vectors respectively. Since z can be reduced to n at most, we can assert that commodity j enters directly or indirectly into the production of commodity i if and only if

$$e_i^{\mathrm{T}}\left[A + A^2 + \cdots + A^n\right]e_j > 0.$$

Commodity j is *basic* if and only if

$$\left[A + A^2 + \cdots A^n\right]e_j > 0. \qquad (1)$$

Let c be the net output vector and x the operation intensity vector. Then

$$x^{\mathrm{T}} = x^{\mathrm{T}}A + c^{\mathrm{T}}$$

i.e.

$$x^{\mathrm{T}} = c^{\mathrm{T}}[I - A]^{-1}$$

Statement (a) asserts that if j is basic then $x^{\mathrm{T}}e_j > 0$ whatever the semipositive vector c is. This is so if and only if

$$[I - A]^{-1}e_j > 0$$

which is a direct consequence of inequality (1) since

$$[I - A]^{-1} = I + A + \cdots + A^m + \qquad (2)$$

The reverse is also true since the vector $A^{n+h}e_j (h \geq 1)$ is a linear combination of vectors $Ae_j, A^2e_j, \ldots, A^ne_j$.

It is easily shown that if and only if all commodities are basic, i.e., if and only if

$$A + A^2 + \cdots A^n > 0$$

matrix A is irreducible, that is, it is not possible to interchange the rows and the corresponding columns to reduce it to the form

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are square submatrices and 0 is a nul submatrix. If some non-basics exist, then matrix A is reducible and as a consequence can be transformed by the same permutation on rows and columns to the following 'canonical' form:

$$A = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ & & & 0 \\ A_{s1} & A_{s2} & \cdots & A_{ss} \end{bmatrix}$$

where $A_{hh} (h = 1, 2, \ldots, s)$ is a square irreducible matrix $[A_{hh} = 0 \text{ only if } \dim(A_{hh}) = 1]$. That is, commodities are partitioned in s groups such that commodities in group $h (h = 2, 3, \ldots, s)$ do not enter either directly or indirectly into the production of commodities in groups $1, \ldots, h-1$. Hence commodities in groups $2, \ldots, s$ are non-basic; commodities in group 1 may or may not be basic.

Sraffa assumes that at least one basic commodity exists. This is equivalent to assuming that

$$A_{21} \geq 0 \tag{3.1}$$

$$(A_{31}, A_{32}) \geq 0 \tag{3.2}$$

$$\left(A_{s1}, A_{s2}, \ldots, A_{s,s-1}\right) \geq 0 \tag{3.s-1}$$

$$\text{if} \dim(A_{11}) = 1 \text{then} A_{11} > 0 \tag{4}$$

Inequality (3 . h − 1), $h = 2, 3, \ldots, s$, asserts that if commodities in group 1 enter directly or indirectly into the production of commodities in groups 1, …, h − 1, then they enter directly or indirectly into the production of commodities in group h (also since A_{hh} is irreducible). If dim $(A_{11}) > 1$, then commodities in group 1 enter directly or indirectly into the production of themselves since A_{11} is irreducible; if dim $(A_{11}) = 1$, then inequality (4) asserts that the commodity in group 1 enters directly into the production of itself.

If p is the price vector, w is the uniform *post factum* wage rate, and r is the uniform rate of profit, then the following equation holds

$$p = (1 + r)Ap + wl, \tag{5}$$

Let us partition vectors p and l in such a way that $p = (p^{\mathrm{T}}{}_1, p^{\mathrm{T}}{}_2, \ldots, p^{\mathrm{T}}{}_s)^{\mathrm{T}}$ and $l = (l^{T}{}_1, l^{T}{}_2, \ldots, l^{T}{}_s)^{T}$, where p_h and l_h are subvectors of the same dimension as A_{hh}. Then eq. (5) can be expanded as

$$p_1 = (1 + r)A_{11}p_1 + wl_1 \tag{6.1}$$

$$p_2 + (1 + r)[A_{21}p_1 + A_{22}p_2] + wl_2 \tag{6.2}$$

$$p_s = (1 + r)[A_{s1}p_1 + A_{s2}p_2 + \ldots + A_{ss}p_s] + wl_2 \tag{6.s}$$

Statement (b) is easily obtained from eq. (6) since inequalities (3) hold. Let

$$R = \sup\left\{\rho \in \mathbb{R} / x \geqq 0, x^{\mathrm{T}} l = 1, x^{\mathrm{T}}[I - (1 + \rho)A] \geqq 0\right\}$$

It is easily recognized that there exists a vector q such that

$$q \geq 0, q^{\mathrm{T}} = (1 + R)q^{\mathrm{T}}A, q^{\mathrm{T}}l = 1. \tag{7}$$

R is called the Standard ratio (see Sraffa 1960, pp. 21 and 26–7). Since the basic commodities are indispensable and non-basics are not [statement (a)], it is easily recognized that the entries of vector q corresponding to non-basics equal zero and the entries of q corresponding to basics are positive. That is, there exists a positive vector q_1 such that

$$q_1^{\mathrm{T}} = (1 + R)q_1^{\mathrm{T}}A_{11}$$

which proves statement (c).

Let

Obviously $R^* \leq R$. Assume that $0 \leq r \leq R^*$,then the theory of M-matrices ensures that matrix $I - (1 + r)A$ is invertible and

$$[I - (1 + r)A]^{-1} = I + (1 + r)A + \ldots + (1 + r)^m A^m + \ldots \geq 0.$$

Moreover,

$$[I - (1 + r)A_{11}]^{-1} > 0$$

because of inequality (1). Thus, if $0 \leq r < R^*$, all prices are non-negative (positive if labour enters directly or indirectly into the production of all commodities, i.e., directly into the production of at least one basic). Assume now that $R^* < R$ and that $R^* \leq r < R$. Then, it is still true that

$$[I - (1 + r)A_{11}]^{-1} > 0$$

but $I - (1 + r)$ does not need to be invertible, and if it is so, its inverse has some negative entries. Thus, statement (d) is proven.

Let z be a semi-positive m-vector, where $m =$ dim (A_{11}). If the numeraire consists of z_1 units of commodity 1, z_2 units of commodity 2, …, z_m units of commodity m, i.e., if

$$\left(z^{\mathrm{T}}, 0^{\mathrm{T}}\right)p = z^{\mathrm{T}}p_1 = 1$$

where 0 is a null vector of dimension $n - m$, then we obtain from eq. (6.1) that

$$w = \begin{cases} \dfrac{1}{z^{\mathrm{T}}[I-(1+r)A_{11}]^{-1}l_1} & \text{if } 0 \le r \le R \\ 0 & \text{if } r = R \end{cases}$$

which proves statement (e).

Sraffa (1960, ch. 4) applies a special *numéraire* because of the useful properties it has. He normalizes prices by setting

$$q^{\mathrm{T}}(I-A)p = 1$$

where q satisfies eq. (7). This *numéraire* is called the *Standard commodity*. Thus, statement (f) is also simply obtained.

The distinction between basics and non-basics is similar to, but different from, the distinction between 'necessary goods' and 'luxury goods'. Such a distinction is mentioned by Adam Smith, who asserted that a tax on one of the latter affects only the price of the taxed commodity; whereas a tax on one of the former affects all prices. Ricardo has remarked that if the real wage is given, a tax on a necessary good affects the profit rate, whereas a tax on a luxury good does not have this property. Dmitriev and Bortkiewicz have clarified that 'necessary goods' must include not only the commodities consumed by workers, but also the commodities which enter directly or indirectly into the production of those commodities (details can be found in Roncaglia 1978).

The analysis by Ricardo, as elaborated by Dmitriev and Bortkiewicz, is definitely correct, if the real wage rate is given. Sraffa suggests that we consider as exogenously given the profit rate, rather than the real wage rate. In this case the distinction between basics and non-basics emerges as the fundamental one.

After the publication of *Production of Commodities*, the distinction between basics and non-basics was the central issue of an exchange of letters between Sraffa and Peter Newman, published as an appendix to an article by K. Bharadwaj (1970). The main issue was the economic rationale of the assumption

$$R^* = R$$

which is necessary and sufficient for positivity of all prices for $0 \le r \le R$.

In Part 2 of *Production of Commodities by Means of Commodities*, Sraffa removes the assumption of single production and allows for joint products and here he provides a general definition of basic and non-basic commodities. Statements (b, i) and (f) are still valid. Statements (a), (b, ii), (c), (d) and (e) do not hold in general. Sraffa's distinction has been formalized by Manara, Steedman, and Pasinetti (see Pasinetti 1980, chs 1, 3, 4).

See Also

- Advances
- Sraffa, Piero (1898–1983)
- Wages in Classical Economics

Bibliography

Bharadwaj, K. 1970. On the maximum number of switches between two production systems. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 106(4).

Flaschel, P. 1982. On two concepts of basic commodities for joint production systems. *Zeitschrift für Nationalömie* 42(3): 259–280.

Pasinetti, L.L., ed. 1980. *Essays on the theory of joint production*. London: Macmillan.

Roncaglia, A. 1978. *Sraffa and the theory of prices*. New York: Wiley.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge, MA: Cambridge University Press.

Basing Point System

M. L. Greenhut

Americans generally consider apple pie, hot dogs, and baseball to be uniquely theirs in origin. Less pride of origin is assigned to the Basing Point System, originally known as Pittsburgh-Plus. However, this lack of pride in origin did not characterize American public opinion during the early years of Pittsburgh-Plus. In fact, at the turn of the century, the members of the Industrial

Basing Point System, Table 1 Sales from Chicago to Pittsburgh, Cleveland, Detroit, and Chicago

	Pittsburgh $	Cleveland $	Detroit $	Chicago $
Base Price	50	50	50	50
Plus Freight from Pittsburgh	0	4	6	8
Delivered Price	50	54	56	8
Minus Freight from Chicago	8	7	5	0
Mill Net	42	47	51	58
Realization at				
Chicago				
Freight	8	3	—	—
Absorption				
Phantom	—	—	1	8
Freight				

*Table taken from Wilcox 1960, p. 269. Its perspective is that of a seller located in Chicago

Commission (the forerunner to the Federal Trade Commission), most members of the Congress of the United States, and later for many, many years the majority of those on the Supreme Court considered the system to be competitive and hence desirable. Before evaluating its pros and cons, a brief inquiry into its early history is in order.

According to testimony given by Henry P. Bope in November 1922 to the Federal Trade Commission, Pittsburgh-Plus was first used in 1880. He told the Commission that steel firms had previously priced their output f.o.b. mill. But beginning in 1880, beams (structural materials) were sold on the basis of the Pittsburgh price plus freight to a given buying point. He testified that this system originated with Carnegie Steel and three small firms, one of which was sited in eastern Pennsylvania, the other two in New Jersey. Having no competitors west of it, Carnegie Steel simply set the Pittsburgh price, and the others used that price plus freight *from Pittsburgh* to all destinations, regardless of the point of shipment. Carnegie Steel thus had access to all markets. The quid pro quo claimed for the small mills was on sales to locations proximate to their plant where they gained a phantom freight, as illustrated in the following hypothetical Table 1. The justification for using Pittsburgh as base was that in those days *most* steel products were shipped from Pittsburgh. But inclusion in the system of tin plates and sheets shortly before the turn of the century is intriguing since there were no such mills in the Pittsburgh area (Fetter 1931, p. 150).

The arguments pro the system (TNEC paper 1940) centred on the claim that the same price exists at a given time in each market with buyers having access to every seller: that is, pure competition. It was further argued that locations of steel plants were due to fundamental economic advantages. These arguments were patently extreme. For example, consider another industry in the United States which also practised base point pricing: the cement industry. Uniquely when Army Engineering opened eleven different bids for cement – each bid was for the same delivered price regardless of point of origin, $3.286854 a barrel (*The Aetna Co. Case* 1946): pure competition? Consider the claim that it is desirable to have many seller alternatives. Under the base point system, each seller (and buyer) foregoes the competitive advantages of proximity. With respect to locational distribution, a small firm subject to basing point pricing can be shown to locate nearer the (base point) production centre than would the same size mill if pricing f.o.b. mill (Greenhut 1970, chapter 7). When the system began to change toward multiple base points – due in part to increased freight rates and challenging litigations in cement, linseed oil, and hardwood flooring – steel industry locations did change considerably. Producers sprang up in the western portion of the United States where they were

able to offer prices lower than the Pittsburgh mills by as much as $3 to $10 a ton (Fetter 1931, p. 160).

What about another argument pro the system, namely that base point pricing prevents formation of local monopolies? This claim, too, is easily revealed to be invalid in theory (Greenhut and Greenhut 1977; Greenhut et al. 1986). It also fails in practice since the prevalence of varying 'net mill' prices, that is, price discrimination, which typifies sales over geographic landscapes, signifies competitive impacts in place of local monopolies (Greenhut et al. 1980; Greenhut 1981).

Phlips (1983, p. 6) defines a *non*-discriminatory price as that which occurs when two varieties of a product are sold by the same seller to two different buyers at the same *net* price, where by net price he means a price corrected for the cost associated with product differentiation. Phlips the claims that discrimination may be as common in the business world as 'it is rare in the economics textbooks' (p. 7). Indeed, it has been emphasized in spatial price theory that discriminatory pricing over geographic space is *the natural pricing* form because spatial markets are naturally separated. In turn, demand elasticities can be expected to differ in each submarket, and different demand elasticities generate different kinds of spatial price discrimination (Hoover 1936–7; Greenhut 1956, p. 157). Considering f.o.b. mill or base point pricing as competitive overlooks these conditions as well as the fact that invasion of markets becomes especially likely when firms discriminate over distances. Not surprisingly, delivered prices of firms in West Germany and Japan were *often* found to be actually lower at greater freight cost distances than at proximate buying sites because of more intense competition at or near rival locations (Greenhut 1981).

Present-day use of the base point system was recently spotlighted by Haddock (1982). By way of examples, Haddock(p. 290) cited the pricing of wheat based on Galveston, Texas, and the pricing of oil. On the intra-national level, cement in Great Britain has been sold under the system. Thus, between January 1982 and September 1983, Dunbar, Aberthaw, Padeswood, Aberdeen and Inverness were used as base points.

How, one may ask, is it possible in nations that consider conspiratorial restraints of trade to be illegal for widely dispersed sellers to quote identical delivered prices for a given product at each market point? The answer is that no specific communication is needed (Machlup 1952, p. 90) *and* for a long time the judiciary in the United States was more concerned with *the means used* than the result itself. If the means did not involve communication, it was not considered a conspiracy (Averitt 1980). However, consider in this regard the recent Southern Plywood Case in the United States.

Douglas Fir, with origins in the state of Washington, had customarily served as the production center for plywood shipments throughout the United States. As many as 100 firms operated 149 mills in the Northwest in 1963, the industry being described as loosely oligopolistic (Loescher 1980, p. 11). In the early 1960s, suppliers of glue developed an adequate bond for manufacturing southern pine plywood; this led to dispersal of the industry.

Following the practice of West Coast producers, the new entrants in the South established a delivered pricing formula which assured the capture of the southern market. They accomplished this by quoting a Portland Base price at a slight differential 'below' the f.o.b. mill price on Douglas Fir plus *West Coast freight to southern destinations*. This signified high mill rates (phantom freight) on sales at southern delivery points close to a mill.

As suggested above, conscious parallelism of action had not typically been considered an unacceptable means. One might thus expect a favourable ruling for Southern Plywood. However, a change has taken place.

Obiter dictum in the *Triangle Conduit and Cable Co v FTC*, 1948, *suggested* that unilateral adoption of basing point systems could be prohibited if the firms were aware others were adopting similar practices. In the Brown Shoe Case (1966), the Federal Trade Commission was empowered to arrest trade restraints in their *incipiency* without proof of outright violation of the antitrust laws.

Now, the argument of the defendants in Southern Plywood was based on an econometric model

for 1967–77 which suggested that Portland base prices varied with housing starts; hence, the defendants said no conspiracy prevails. But, as noted in the plaintiff's successful charge to the jury, individual self-interest would permit continuance of the system if and only if each expected the others '. . . to continue the common practice' (p. 330). Thus, Loescher (1980, p. 29), citing Turner (1962), Stigler (1949), Posner (1976) and himself (1959), notes that all of these writers are of the belief that *inferring a civil conspiracy* behind a basing point system is today intrinsic to American antitrust policy and practice. To assert next that the Basing Point system is therefore on its way out of use in the United States would, however, be a stronger inference than one can soberly make. This is especially so given the freight rate zones and basing lines of American railroads that sellers can avail themselves of in establishing difficult to evaluate variants of a *multiple* base point system. The prevalence of the system elsewhere in the world would also appear to depend on whether conspiracies and restraints of trade are strongly condemned *and* whether conscious parallelism of action can be identified and then inferred as a conspiracy.

See Also

- ▶ Location of Economic Activity
- ▶ Price Discrimination

Bibliography

Aetna Portland Cement Co. v FTC. (1946). 157 F. 2nd 533 (1946), Respondents Brief p. 127.

Averitt, N.W. 1980. The meaning of 'unfair' methods of competition in section 5 of the Federal trade Commission Act. *Boston College Law Review* 21(2): 227–300.

Brown Shoe Co. (1966). 384 US 316.

Federal Trade Commission. 1922. Docket 760, Pittsburgh-Plus Complaint, Record pp. 10861–2.

Fetter, F.A. 1931. *The masquerade of monopoly*. New York: Harcourt/Brace & Co.

Greenhut, M.L. 1956. *Plant location in theory and practice*. Chapel Hill: University of North Carolina Press; 4th printing, Westport, Conn.: The Greenwood Press, 1983.

Greenhut, M.L. 1970. *A theory of the firm in economic space*. New York: Appleton-Century-Crofts. 2nd printing, Austin, Texas: Lone Star Publishing Co., 1974.

Greenhut, M.L. 1981. Spatial pricing in the USA, West Germany, and Japan. *Economica* 48(189): 79–86.

Greenhut, M.L., and J. Greenhut. 1977. Nonlinearity of delivered price schedules and predatory pricing. *Econometrica* 45(8): 1871–1875.

Greenhut, M.L., J. Greenhut, and S. Li. 1980. Spatial pricing patterns in the United States. *Quarterly Journal of Economics* 94: 329–350.

Greenhut, M.L., G. Norman, and G. Hung. 1986. *Imperfect competition: A spatial approach*. Cambridge: Cambridge University Press.

Haddock, D.O. 1982. Basing-point pricing: Competitive vs. collusive theories. *American Economic Review* 72(3): 289–306.

Hoover, E.M. 1936–7. Spatial price discrimination. *Review of Economic Studies* 4: 182–91.

In re *Plywood Antitrust Litigation*. (1978). MOL Docket Number 159, United States District Court, Eastern District of Louisiana.

Loescher, S.M. 1959. *Imperfect collusion in the cement industry*, 18–22. Cambridge, MA: Harvard University Press, 232–40.

Loescher, S.M. 1980. Economic collusion, civil conspiracy, and treble damage deterrents: The Sherman act breakthrough with southern plywood. *Quarterly Review of Economics and Business* 20(4): 6–35.

Machlup, F. 1952. *The political economy of monopoly*. Baltimore: Johns Hopkins Press.

Phlips, L. 1983. *The economics of price discrimination*. Cambridge: Cambridge University Press.

Posner, R.A. 1976. *Antitrust law*. Chicago: University of Chicago Press.

Stigler, G.J. 1949. A theory of delivered prices. *American Economic Review* 39: 1143–1159.

TNEC Papers. 1940. *The basing point method*, vol. 3. Published by The Corporation.

Triangle Conduit & Cable Co. v FTC. (1948). 168 F 2nd 1975 (7th Cir. 1948).

Turner, D.F. 1962. The definition of agreement under the Sherman Act. *Harvard Law Review* 75: 655.

Wilcox, C. 1960. *Public policies toward business*. Revised edn, Homewood: Richard D. Irwin.

Bastable, Charles Francis (1855–1945)

John A. Bristow

Born in Co. Cork in 1855, Bastable graduated in history and political science from Trinity College, Dublin in 1878 and was called to the Irish Bar in 1881. The next year was to be the first of his

fifty-year tenure of one of the oldest chairs of political economy in the British Isles – the Whately Chair at Trinity College. Throughout the whole of that period he also occupied a succession of chairs in legal subjects, including the Regius Chair of Laws at Trinity (1908–1932). He was a member of the first Council of the Royal Economic Society and among his scholastic honours were his presidency of Section F of the British Association in 1894 and his election as a Fellow of the British Academy in 1921. He died in Dublin in 1945.

Bastable's place in the history of economics is as an expositor of and commentator on classical doctrines, rather than as an original thinker. His main areas of interest were, first, trade and commercial policy and, second, public finance. His *The Commerce of Nations* is a spirited, semi-popular defence of free trade. Whilst written as a tract for the times, parts of this book have a continuing relevance and appeal, most notably in a chapter entitled 'Economic Arguments for Protection', which remained unchanged through successive editions over thirty years and which, while obviously tendentious, effectively rebuts some of the cruder fallacies which are still paraded today.

Bastable's concern for the pure and monetary theory of trade found expression in several journal papers (notably in the *Economic Journal*) and in his treatise, *The Theory of International Trade*. The latter is firmly in the English classical tradition, the discussion of comparative costs, for example, being essentially an exposition of Mill, defending him against such critics as Cournot. Bastable's largest single work, *Public Finance*, was written explicitly as a textbook and, for its scope and clarity, deserves an honoured place in any history of that *genre*. The theory is, again, English classical, but the work as a whole is impressively eclectic, covering expenditure, taxation and debt with a wealth of institutional detail and juxtaposing arguments and examples from a large range of European and American sources.

In his writings at least, Bastable never appeared to recognize the significance of neoclassical innovations. He cites Marshall's *Principles* in various works, but not in a context which suggests that anything of analytical significance is contained therein. Indeed, he explicitly rejects as 'unsuitable' to the treatment of tax incidence the unity of value and distribution theory which he properly describes as 'the whole tendency of modern economic science' (*Public Finance*, 2nd edn, p. 331).

Perhaps the best indication of his doctrinal and methodological position is to be found in his 1894 presidential address to the British Association. He praises the German historical school, stresses the importance of sociology to the economist and urges the integration of economics with 'political science, jurisprudence and the scientific principles of administration'. This view that economics is part of the seamless garment of the social sciences informed all the work of this humane and scholarly man.

Selected Works

1887. *The theory of international trade*, 2nd–4th eds. Dublin. London: Macmillan, 1897–1903.
1891. *The commerce of nations*, 9th ed. London: Methuen, 1923.
1892. *Public finance,* 2nd ed., 1895; 3rd ed. London: Macmillan, 1903.

Bastard Keynesianism

G. C. Harcourt

This is the name given by Joan Robinson to certain developments which occurred in Keynesian economics following the publication of the General Theory, principally in the USA. They culminated in the system of thought which is more usually known as the neoclassical synthesis. The basic idea was that the notion of equilibrium of the economic system in traditional theory (traditional in the sense of Harrod 1937), in which all markets (including the labour market) clear, was an

accurate description of the outcome of tendencies in the economic system. However, the forces in the economy which allowed this position to be sought were weak, were often frustrated by rigidities and imperfections, and in any event took a long time to work themselves out. The economy would often be found for long periods of time experiencing sustained unemployment due to a deficiency of overall demand. Therefore there was a role for government intervention to reinforce and speed up the processes whereby the economy found its way to its full employment equilibrium position; once there, the traditional theory of resource allocation and income distribution would come into its own again. That is to say, an equilibrium position has been shown to exist within the bounds of traditional theory, so that Keynes has a place not so much as a theorist but as a sensible propagator and rationalizer of policies in the short period, over the cycle and perhaps permanently, as the average level of unemployment reflected a permanent tendency to a deficiency in aggregate demand.

The starting point of this analysis was the expression of what was argued to be the analytical core of the *General Theory* in terms of the IS/LM general-equilibrium framework, associated especially with Hicks (although both Harrod 1937; Meade 1937, wrote down the same system in their interpretative papers of the *General Theory* in the same year – 1936 – as Hicks 1937). The attempt to confine Keynes's contributions within a small general equilibrium model allowed the neoclassical synthesis to occur. For, following the contributions of Patinkin in the 1950s, recognition of the existence and role of the real balances or Pigou effect made it possible to define a full employment equilibrium: that is, to prove existence in the sense that, given the supply of money and that the key relationships of the system were stable and especially that their positions were independent of the processes and paths by which the equilibrium was found, there must always be a value of the general price level which implied a level of aggregate demand that was consistent with full employment. There was the additional proviso that as long as this position had *not* been attained, the general price level could not be constant because of competitive pressures on the money-wage rate in the labour market. This led to the interpretation of Keynes as the economics of dis-equilibrium, a situation which happened to be the usual state of the real world unless the government acted but which was not *theoretically* that interesting.

This interpretation of Keynes's contributions was regarded by Joan Robinson in particular (but also by Kahn, Kalecki and Shackle amongst others; for a contemporary view, see Chick 1983) as illegitimate – hence the name, bastard Keynesianism. (Joan Robinson coined the phrase, 'the bastard Keynesians', in 1962 in her review of Harry Johnson's *Money, Trade and Economic Growth* (1962). She took particular exception to his assessment of the *General Theory* 25 years after its publication, arguing that what Johnson and other 'bastards' of his generation saw as weaknesses were in fact strengths – to wit, a sense of time, of the structure of society and of economic life as a process.) But for Joan Robinson and other kindred souls, Keynes *had* established, through his theories of investment behaviour and the consumption function, that there was no automatic tendency for the economy to gravitate towards a full-employment equilibrium. Rather, there was an underemployment position (the interpretation of the characteristics of which varied according to whether it was Joan Robinson and her followers or Garegnani (1978, 1979), Eatwell (1979) and Milgate (1982) and their followers who described it). Keynes argued that, because of the uncertainty which of necessity must surround decisions about investment and holding money, and because producers in a monetary production economy of necessity must produce in anticipation of demand and of a money profit, and must make contracts in money terms, there are no necessary equilibrating forces which take the economy to full employment either at a point in time or over the cycle.

Moreover, Keynes himself stressed both the likely instability of his core functions, especially the investment and liquidity preference functions, *and* the dependence of this instability on movements in the economy itself, so that positions were

not independent of paths. The IS/LM apparatus was therefore peculiarly unsuited to capture this vision of the operation of the economy, and the neoclassical synthesis itself was a denial of the revolution both in vision and in method which Keynes had provided. Furthermore, just because the neoclassical synthesis version dominated the profession when the monetarist counter-revolution came to prominence in the late 1960s and early 1970s, Keynesians were weakened in their fight back because they had already, unnecessarily and illegitimately, conceded the framework of the approach within which the battle was to be fought.

See Also

► Robinson, Joan Violet (1903–1983)

Bibliography

Chick, V. 1983. *Macroeconomics after Keynes, a reconsideration of the general theory*. Oxford: Phillip Allan.

Eatwell, J. 1979. *Theories of value, output and employment. Thames papers in political economy*. London: Thames Polytechnic.

Garegnani, P. 1978. Notes on consumption, investment and effective demand: I. *Cambridge Journal of Economics* 2: 335–354.

Garegnani, P. 1979. Notes on consumption, investment and effective demand: II. *Cambridge Journal of Economics* 3: 63–82.

Harrod, R.F. 1937. Mr Keynes and traditional theory. *Econometrica* 5: 74–86.

Hicks, J.R. 1937. Mr Keynes and the 'Classics'; a suggested interpretation. *Econometrica* 5: 147–159.

Johnson, H.G. 1962. *Money, trade and economic growth*. London: Allen & Unwin.

Kalecki, M. 1936. Pare uwag o teorii Keynesa. *Economista* 3, reprinted in M. Kalecki, *Kapitalizm, Koniunktura i Zatrudmienie*, Warsaw: PWN, 1979. Trans. by F. Targetti and B. Kinda–Hass as 'Kalecki's review of Keynes' *General Theory', Australian Economic Papers* 21, December, 244–60.

Meade, J.E. 1937. A simplified model of Mr Keynes' system. *Review of Economic Studies* 4: 98–107.

Milgate, M. 1982. *Capital and employment. A study of Keynes's economics*. London: Academic.

Robinson, J. 1962. Review of H.G. Johnson, *Money, trade and economic growth*, 1962. *Economic Journal* 72: 690–92. Reprinted in J. Robinson, *Collected economic papers*, Vol. 3. Oxford: Basil Blackwell, 1965.

Bastiat, Claude Frédéric (1801–1850)

R. F. Hébert

Keywords

Bastiat, C. F.; Cobden, R.; Dupuit, A.-J.-E. J.; Free trade; Harmonism; Service theory of value

JEL Classifications

B31

French economist and publicist, born at Bayonne on 30 June 1801, the son of a merchant in the Spanish trade; died in Italy, at Rome, on 24 December 1850. Orphaned at the age of nine, Bastiat nevertheless received an encyclopedic education before entering his uncle's business firm in 1818. By 1824 he was expressing dissatisfaction with his employment. Upon inheriting his grandfather's estate in 1825, he left business and became a gentleman farmer at Mugron, but showed no more aptitude for agriculture than he had for commerce. So he became a provincial scholar, establishing a discussion group in his village and reading voraciously. His later writings show familiarity with the works of French, British, American and Italian authors, among them Say, Smith, Quesnay, Turgot, Ricardo, Mill, Bentham, Senior, Franklin, H.C. Carey, Custodi, Donato and Scialoja.

Bastiat left France in 1840 to study in Spain and in Portugal, where he tried unsuccessfully to establish an insurance company. Returning to Mugron, he learned (in the course of seeking information for his study club) of Cobden's Anti-Corn Law League and became an ardent free-trader (the 'French Cobden'). As a complete unknown in economics, he submitted a stirring article to the *Journal des économistes* in 1844, dealing with the influence of protectionism on France and England. It created an immediate sensation and raised a clamour for more from the editors. This response encouraged Bastiat's *Economic Sophisms,* which quickly sold out upon its

publication in 1845, and was soon thereafter translated into English and Italian. In 1846 Bastiat moved to Paris, where he established the Association for Free Trade and quickened his literary activity, endangering his frail health in the process. A torrent of articles, pamphlets and books now flowed from his talented pen, undoubtedly made possible in such short order by the preceding 20 years of practically uninterrupted reflection. Some scholars say the frenzy produced more heat than light, yet on the whole, economics is better off for Bastiat's Herculean efforts.

Bastiat was one of several writers (Quesnay, Smith, Say and Carey were the others) who formed the doctrine of Harmonism, or the optimistic idea that class interests naturally and inevitably coincide so as to promote economic development. The major challenge to this view came from Ricardo and Malthus, whose theories cast a sinister shadow over the prospect of economic progress. As against Ricardo's system, Bastiat erected a theory of value based on the idea of service. He distinguished between utility and service, identifying the former as insufficient, of itself, to establish value, because certain free goods (sun, air, water) have utility. Bastiat considered all commercial transactions as exchanges of service, with value measured in terms of the trouble a buyer saves by making the purchase.

J.E. Cairnes complained that this merely confounded what Ricardo had sought to delineate, namely those cases in which value is proportioned to effort and sacrifice from those in which it is not. A more fundamental criticism is that Bastiat's theory, notwithstanding denials to the contrary, is simply a labour theory in different guise. It is noteworthy, however, that Bastiat's idea bears a close resemblance to the notion of 'public utility' which Dupuit applied so successfully to the measure of gain from transport improvements, and in which reduction of costs effected by the improved service became the central issue. Yet any connection between the two, tenuous as it may be, must be considered to run from Dupuit to Bastiat rather than the reverse, since Dupuit published his famous article on public works and marginal utility before Bastiat abandoned his earlier polemics in favour of more 'constructive' attempts at theory. Bastiat's theory of rent, also clearly aimed against Ricardo, denied the notion of unearned income, again advancing the view that the value of land (always in the absence of government interference) derives entirely from the services it renders.

Generally, judgement on Bastiat has been that he made no original contributions to economic analysis. Cairnes, Sidgwick and Bohm-Bawerk discounted his pure economics completely. Marshall said that he understood economics hardly better than the socialists against whom he declaimed. And Schumpeter declared that Bastiat was not a *bad* theorist, he was simply no theorist at all.

Schumpeter also described Bastiat as 'the most brilliant economic journalist who ever lived', and so weighty a thinker as Edgeworth praised Bastiat's genius for popularizing, in the best sense of the term, the economic discoveries of his predecessors. Almost all commentators agree that Bastiat was unrivalled at exposing economic fallacies wherever he found them, and he found them everywhere. He was quite simply a genius of wit and satire, frequently described as a combination of Voltaire and Franklin. He had the habit of exposing even the most complex economic principles in amusing parables that both charmed and educated his readers. His writings retain their currency, even today. And as Hayek has reminded us in his introduction to Bastiat's *Selected Essays,* his central idea continues to command attention: the notion that if we judge economic policy solely by its immediate and superficial effects, we shall not only not achieve the good results intended, but certainly and progressively undermine liberty, thereby preventing more good than we can ever hope to achieve through conscious design. This principle is exceedingly difficult to elaborate in all of its profundity, but it is one which has galvanized the thought of contemporary economists, Hayek and Friedman.

Over the long haul, Bastiat's influence has waxed and waned. In his own day he received the ready support of Dunoyer, Blanqui, Chevalier and Garnier. Francis A. Walker introduced his doctrines into America at about the time of the Civil War. Pre-First World War French liberals such as Leroy-Beaulieu, Molinari and Guyot relied on his authority. Bastiat's ideas subsequently went into a long decline, only to become resurgent in the late 20th

century among libertarian economists dissatisfied with Keynesian orthodoxy and Marxist alternatives. Ironically, Bastiat's originality is exhibited most in his contribution to political theory, which has drawn surprisingly little attention to this day.

Selected Works

1844. De l'influence des tarifs franeais et anglais sur l'avenir des deux peuples. *Journal des économistes* 9: 244–271.
1964a. *Economic sophisms*. Trans. A. Goddard. Princeton: D. Van Nostrand.
1964b. *Selected essays on political economy*. Trans. S. Caine. Princeton: D. Van Nostrand.
1964c. *Economic harmonies*. Trans. W.H. Boyers. Princeton: D. Van Nostrand.

Bibliography

Baudin, L. 1962. *Frédéric Bastiat*. Paris: Dalloz.
Hayek, F.A. 1964. Introduction to F. Bastiat. In *Selected essays on political economy*. Princeton: D. Van Nostrand.
Russell, D. 1963. *Frédéric Bastiat: Ideas and influence*. Irvington-on-Hudson: Foundation for Economic Education.

Baudeau, Nicolas (1730–*c*1792)

Peter Groenewegen

Keywords
Baudeau, N.; Free trade; Gournay, Marquis de; Hoarding; Luxury; Mirabeau, V. R., Marquis de; Monopoly; Net product; Physiocracy; Productive vs. unproductive expenditure; Sumptuary laws; Surplus

JEL Classifications
B31

Born at Amoise, Baudeau entered the church, becoming a canon and professor of theology at the Chancelade Abbey. He was subsequently called to Paris in the service of Archbishop de Beaumont. In 1765, Baudeau founded the periodical *Ephémérides,* becoming its first editor till late 1768 and again during its two subsequent revivals. Converted to Physiocracy by Mirabeau in 1768, he became one of its most active propagandists through the many articles, pamphlets and books he produced. He died insane in Paris *circa* 1792 (Coquelin and Guillaumin 1854, I, p. 148). Daire (1846, pp. 652–4) provides a bibliography of the economic writings and reprints his long introduction to economic philosophy (Baudeau 1771) and his explanations of the *Tableau économique* (Baudeau 1767–8), which Marx (1962, p. 324) found helpful for clarifying some of its more difficult points and which remains a most useful introduction to Physiocracy and the *Tableau's* intricacies. Baudeau (1771) is noteworthy for its concise definition of monopoly as 'everything which by force limits the numbers and competition of buyers and sellers' (p. 327) and its direct attribution to Gournay of the phrase, *laissez les faire* (p. 323).

Perhaps the most interesting of Baudeau's many writings is his systematic exposition and development of the Physiocratic theory of luxury (Baudeau 1767), the most complete version of that doctrine and as such wrongly ignored (Dubois 1912, pp. v–vi). Inspired by the Swedish sumptuary laws of 1767, and bearing in mind the Physiocratic division of output between necessary expenses and disposable net product, the essay clearly defines luxury as 'that subversion of the natural and essential order of national expenditure which increases the total of unproductive expenditure to the detriment of that which is used in production and at the same time to the detriment of production itself' (Baudeau 1767, p. 14). In other words, disposal of the net product when in direct agricultural investment or in spending which directly or indirectly enhances the demand for agricultural produce is productive: other uses of the surplus are wasteful, luxury spending. For example, hoarding which detracts from demand for agricultural produce, is luxury; importing commodities from abroad, if this increases overseas demand for domestic produce and thereby augments productive

expenses, is not. Sumptuary laws are therefore not appropriate for curtailing luxury; free trade and a more simple pattern of consumption channelling more demand to the agricultural sector, are much more effective. In short, ostentation in consumption is to be preferred to ostentation in display and ornament, since the former creates a greater market for agricultural produce and hence for all production. As Meek (1962, p. 318) points out, this 'theory of luxury, with its distinction between productive and unproductive expenditure out of revenue, was much more useful to Smith and Ricardo than it was to the underconsumptionists', despite its emphasis on consumption spending as a factor in stimulating production.

See Also

► Ephémérides du citoyen ou chronique de l'esprit National

Selected Works

1767. Principes de la science morale et politique sur le luxe et les lois somptuaires. *Ephémérides* 1(1), January, and 1(3). Reprinted, ed. A. Dubois. Paris: Marcel Rivière, 1912.

1767–8. *Explication du Tableau économique à Madame de* ***. Paris: Delalain, 1776.

1771. *Première introduction à la Philosophie économique ou analyse des états policés*. In *Physiocrates*, ed. E. Daire. Paris: Guillaumin, 1846.

Bibliography

Coquelin, Ch., and M. Guillaumin. 1854. Baudeau. In *Dictionnaire de l'économie politique*. Paris: Guillaumin.

Daire, E. 1846. Notice sur la vie et les travaux de l'Abbé Baudeau. In *Physiocrates*, ed. E. Daire. Paris: Guillaumin.

Dubois, A., ed. 1912. *Nicolas Baudeau, Principes de la science morale et politique*. Paris: Marcel Rivière.

Marx, K.H. 1962. *Theories of surplus value,* Part I. Moscow: Foreign Languages Publishing House.

Meek, R.L. 1962. *The economics of physiocracy.* London: Allen & Unwin.

Bauer, Otto (1881–1938)

Tom Bottomore

Born 5 September 1881, Vienna; died 4 July 1938, Paris. A member of a talented Jewish family and the only son of a textile manufacturer, Bauer became interested in Marxism and the 'revisionist' controversy while still in high school, and went on to study philosophy, law and political economy at the University of Vienna. He became the leader of the Austrian socialist party (SPÖ) and a prolific writer on economic and political questions. Bauer is best known for his study of nationalities and nationalism (1907), which remains the classic Marxist work on the subject, but he also wrote extensively on economics and his first major essay (1904), which brought him to the notice of Karl Kautsky, discussed the Marxist theory of economic crises. In his early writings he adopted a 'disproportionality' theory such as Hilferding expounded more fully in *Finance Capital* (1910); that is, a theory which sees the fundamental causes of crises in the 'anarchy of capitalist production', and particularly in the disproportion which regularly emerges between production in the two sectors of capital goods and consumer goods. However, in his last published book (1936) he propounded an underconsumption theory of crises which subsequently influenced the work of Sweezy. In the course of his analyses of economic crises Bauer introduced, or emphasized more strongly than other Marxist writers, such factors as the existing stock of capital, technical progress, and population growth.

Bauer also discussed economic questions in a broader context in his study of the development of capitalism and socialism after World War I, of which only the first volume was published (1931). In this work he examined the rationalization of capitalist production in three spheres: technical rationalization, the rationalization and intensification of work, and the rationalization of the enterprise (especially the growth of 'scientific management'). The final part of the book dealt with the limits to capitalist rationalization

revealed by the economic crisis, its consequences for the working class, which he analysed in terms of a distinction between the 'labour process' (a concept which has become central in much recent Marxist political economy) and the 'life process', and the nature of rationalization in a socialist society.

Besides his major studies of nationalism and of the capitalist economy Bauer published many other important essays and books: on the Austrian revolution (where he strongly opposed the idea of a Bolshevik type revolution and began to elaborate his conception of the 'slow revolution'), on violence in politics and the doctrine of 'defensive violence', on fascism, on the philosophical foundations of Austro-Marxism, and on Marxism and ethics. His work as a whole represents one of the most important and interesting contributions to Marxist thought in the 20th century. The defeat of the SPÖ in the civil war of 1934, which drove Bauer into exile, was attributed by some critics to his excessively cautious and gradualist policies; on the other hand, the social, educational and cultural achievements of 'Red Vienna' in the 1920s and early 1930s showed the effectiveness of such policies when the socialists were in power, and they have had a major influence on Austria's development since 1945.

Selected Works

1904–5. Marx' Theorie der Wirtschaftskrisen. *Die Neue Zeit* 23.
1907. *Die Nationalitätenfrage und die Sozialdemokratie*. Vienna: Wiener Volksbuchhandlung. 2nd enlarged edition with new Preface, 1924.
1923. *Die Österreichische Revolution*. Abridged English version. New York: Burt Franklin, 1925; reprinted, 1970.
1931. Kapitalismus und Sozialismus nach dem Weltkrieg. Vol. I. *Rationalisierung oder Fehlrationalisierung?* Vienna: Wiener Volksbuchhandlung.
1936. *Zwischen zwei Weltkriegen?* Bratislava: Eugen Prager Verlag.

Bibliography

Bottomore, T., and P. Goode, eds. 1978. *Austro-Marxism*. Oxford: Clarendon Press.
Botz, G. 1974. Genesis und Inhalt der Faschismustheorien Otto Bauers. *International Review of Social History* 19: 28–53.
Braunthal, J. 1961. *Otto Bauer: Eine Auswahl aus seinem Lebenswerk*. Vienna: Wiener Volksbuchhandlung.

Bauer, Peter Thomas (1915–2002)

James A. Dorn

Keywords
Bauer, P.T.; Capital accumulation; Central planning; Democracy; Development economics; Economic freedom; Foreign aid; Free trade; Freedom of contract; Gains from trade; Institutional economics; Limited government; Poverty; Poverty traps; Property rights; Total factor productivity

JEL Classifications
B31

Peter (Lord) Bauer, one of the pioneers of early post-Second World War development economics, stood almost alone in the 1940s and 1950s in questioning the prevailing orthodoxy.

Born in Budapest on 6 November 1915, he was the son of a bookmaker. Bauer left Hungary in 1934 to study at Cambridge University, where he earned a first-class degree in economics from Gonville and Caius College in 1937. He returned home to complete his law degree at Budapest University, and then took a job in London with the trading firm of Guthrie & Company. In 1947 he was appointed a lecturer in agricultural economics at London University. From 1948 to 1956 he was a lecturer in economics at Cambridge University, and then became Smuts Reader in Commonwealth Studies. In 1960 Bauer accepted a professorship at the London

School of Economics, and took emeritus status in 1983. Prime Minister Margaret Thatcher elevated Bauer to the House of Lords, as a life peer, in 1982. Lord Bauer was a fellow of the British Academy and of Gonville and Caius College. He was the first recipient of the Milton Friedman Prize for Advancing Liberty, a $500,000 prize awarded every two years by the Cato Institute. The award cited Bauer's 'tireless and pioneering scholarly contributions to understanding the role of property and free markets in wealth creation'. Peter Bauer died on 2 May 2002 at the age of 86.

In the early post-war era, orthodox development economists held that there was a 'vicious circle of poverty'. They assumed that low incomes in less developed countries would prevent sufficient domestic saving and capital accumulation, which were seen as essential for growth. Moreover, poor people were assumed to be incapable of readily responding to market incentives or to have the foresight to save and invest, investment opportunities were seen as narrowly limited, and external trade was viewed as ineffective or even harmful. Poverty was therefore regarded as self-perpetuating. The only escape was to generate a 'big push' by comprehensive central planning and by relying on external assistance.

Bauer's first-hand observations during his extensive work in south-east Asia and in British West Africa in the 1940s and 1950s led him to question the conventional wisdom. In his classic studies of the rubber industry in Malaya (Bauer 1948) and small traders in West Africa (Bauer 1954), he found strong evidence that poor people can lift themselves out of poverty by hard work, entrepreneurial activities, and internal and external trade – provided they have the freedom to do so. He was fond of saying, 'If the notion of the vicious circle of poverty were valid, mankind would still be living in the Old Stone Age'.

Rather than advocate a state-led development model, which was in high fashion at the time, Bauer argued that investment planning, compulsory saving, protectionist trade policies, marketing boards, and government-to-government transfers (foreign aid) would politicize economic life, empower the ruling class, and perpetuate poverty. His views have been vindicated by the failure of comprehensive economic planning and by the ineffectiveness of official aid to spur development.

For Bauer, the essence of economic development is to increase 'the range of effective alternatives open to people' – that is, to increase economic freedom. Until recently, this classical-liberal view was largely invisible. Bauer was among the first to downplay the importance of physical capital accumulation as a precondition for growth. His focus was on institutions and incentives, and especially on the dynamic gains from trade. Total factor productivity is a black box that must be opened to understand the underlying forces of the development process. Bauer was sceptical that those forces could be precisely modelled or that there could be a general theory of development. The process was much too complex.

The primary role of government, in Bauer's view, is to protect private property rights and freedom of contract so that individuals are free to choose and to trade. Conditions will then be conducive to develop and to prosper. Limited government is more important than democracy, in this respect. Hong Kong has few natural resources but has limited government and free trade, and was able to escape the 'poverty trap' – without comprehensive planning or foreign aid.

Bauer, like Ronald Coase, relied on direct observation, an understanding of institutions and history, and sound economic logic to overturn conventional wisdom. When nearly everyone was focusing on capital accumulation as the primary determinant of growth, Bauer (1957a, p. 119) argued, 'It is more meaningful to say that capital is created in the process of development, rather than that development is a function of capital'.

In his final book, *From Subsistence to Exchange and Other Essays* (2000), Bauer summarized his market-liberal vision of the development process:

- 'Economic performance depends on personal, cultural, and political factors, on people's

attitudes, motivations, and social and political institutions.'
- 'Contacts through traders and trade are prime agents in the spread of new ideas, modes of behavior, and methods of production.'
- 'Development aid is thus clearly not necessary to rescue poor societies from a vicious circle of poverty. Indeed, it is far more likely to keep them in that state.'

Those ideas were controversial for many years, but are now more readily accepted in the field of development economics. Bauer deserves much credit for that reversal.

See Also

► Growth and Institutions

Selected Works

1948. *The rubber industry.* London: Longmans, Green and Co.
1954. *West African trade.* Cambridge: Cambridge University Press.
1957a. *Economic analysis and policy in underdeveloped countries.* Durham: Duke University Press.
1957b. (With B.S. Yamey.) *The economics of underdeveloped countries.* Chicago: University of Chicago Press.
1976. *Dissent on development*, revised ed. Cambridge, MA: Harvard University Press.
1991. *The development frontier.* Cambridge, MA: Harvard University Press.
2000. *From subsistence to exchange and other essays.* Princeton: Princeton University Press.

Bibliography

Dorn, J.A. 2002. Economic development and freedom: The legacy of Peter Bauer. *Cato Journal* 22: 355–371.
Yamey, B.S. 2005. Peter Bauer: An unusual applied economist. *Cato Journal* 25: 449–453.

Baumol's Cost Disease

The Tendency for Costs and Prices to Rise in Sectors That Cannot Easily Incorporate Technological Advances, Relative to Technology-Adopting Sectors

Charles M. Gray
University of St. Thomas, Minneapolis, Minnesota, USA

Abstract

The tendency for costs and prices to rise in sectors that cannot easily incorporate technological advances, relative to technology-adopting sectors.

The so-called cost disease was initially diagnosed by William Baumol and William Bowen (Baumol, W.J., and W. Bowen. 1966. *Performing arts: The economic dilemma*. New York: Twentieth Century Fund.) in their mid-1960s study of the performing arts on behalf of the Ford Foundation. Their observations regarding differential productivity enhancements in the "progressive" and "non-progressive" or "stagnant" sectors helped to explain the earnings gap in the arts as well as elements of urban crises and rising costs in many service sectors. Many theoretical and empirical studies later, the concept remains contentious, with supporters and doubters still.

Keywords

Productivity; Unbalanced growth; Cost disease

JEL Categories

D2; H7; I1; I2; J31; L3; O3; O4; Z11

Introduction

In the mid-1960s, the arts world was both shaken and emboldened when a research study by William Baumol and William Bowen explained an underlying cause of the performing arts earnings

gap, the growing distance between operating expenses, and the ability to meet these expenses with ticket sales and other earned revenues. The shortfall required some combination of government subsidies and private donations if arts organizations were to continue to survive.

The performing arts are a prime example of an industrial sector that cannot easily incorporate technological advances into its production process. Baumol subsequently extended this analysis to additional industries, principally education and health care (Baumol 1996, 2012). Cultural economists delved often and deeply into the issue (Flanagan 2012; Towse 1997), and additional research has uncovered evidence of the disease in other industries and other countries (Bates and Santerre 2015; Hartwig 2008; Last and Wetzel 2011).

A number of subsequent writers have contended that the disease *can be* "cured" (Brooks 1997), that it *has been* cured (Triplett and Bosworth 2003; Bosworth and Triplett 2007; Gordon 2016), or that they "do not believe" in the disease (Cowen 1997). The cure seems, however, not to have been permanent, as the disease keeps reappearing, most recently gaining renewed interest as related matters have become politicized.

Interindustry differences in the trend of productivity growth have one very important consequence: They cause related but opposite differences in the trend of unit production costs. The cost of services in which output per work-hour increases slowly rises relative to the cost of goods for which gains in output per work-hour are more rapid, and the cost of services such as education or the live performing arts, in which output per work-hour is almost unchangeable, rises most of all.

In the remainder of this entry, we explore the productivity lag problem, illustrate the resulting cost impact with specific reference to the live performing arts, consider the possibility of cures to the cost disease, and draw some conclusions regarding future impact.

The Productivity Lag Argument

The argument posed in this entry, adapted from Baumol and Bowen's original analysis, can be summarized as follows: Costs in those industries that are resistant to productivity increases will rise relative to costs in the economy as a whole because wage increases in those industries have to keep up with those in the general economy even though productivity improvements lag behind.[1] All industries compete to hire workers in an integrated labour market, and the wages in stagnant industries must therefore rise over time by the same proportion as wages in the general economy if those industries are to remain viable.

In any economy or sector there are five possible sources of growth in output per work-hour:

1. Increased capital per worker. If workers are provided with – or replaced by – more machinery, output per work-hour rises: ten workers with two front-loaders and two trucks can move more earth in an hour than ten workers with one front-loader and one truck.
2. Improved technology. Technology can be defined as the state of knowledge about methods of production. The introduction of assembly line robotics vastly increased output per work-hour in manufacturing.
3. Increased labour skill. Obviously, if workers are more skillful, they can produce more output per hour. Skills may be improved by either education or on-the-job training.
4. Better management. If managers develop more efficient ways of organizing the production process, output per work-hour will rise.
5. Economies of scale. In some production processes, such as automobiles, output per unit of input rises when the scale of production increases. Such industries are said to enjoy economies of scale and, among other things, display increased output per work-hour as the scale of output rises.

As one might guess from this list of causes, productivity increases are achieved most readily in industries that make use of a lot of productive

[1]Portions of this entry borrow from and build upon the corresponding discussion in Gray, Borowiecki, and Heilbrun (forthcoming).

equipment. Output per worker can then be increased either by using more machinery or by investing in new equipment that embodies improved technology. As a result, in the typical manufacturing industry the amount of labour time needed to produce a physical unit of output declines dramatically decade after decade. The service industries are at the other end of the spectrum. Machinery, equipment, and technology play only a small role and, in any case, change very little over time.

That is not to say that technological improvements are entirely absent from the arts and other service industries. For example, stage lighting has been revolutionized by the development of electronic controls and audience comfort greatly enhanced by air-conditioning, which also facilitates longer seasons and more flexible scheduling. But these improvements are not central to the business at hand. As Baumol and Bowen (1966, 164) pointed out, the conditions of production themselves preclude any substantial change in productivity because "the work of the performer is an end in itself, not a means for the production of some good." Since the performer's labour is the output – the singer singing, the dancer dancing, the pianist playing – there is really no way to increase output per hour. It takes four musicians as much playing time to perform a Beethoven string quartet today as it did when it was first performed over 200 years ago.

Of the five sources of increased productivity cited, only economies of scale, in this case the result of longer seasons, is really effective in the live performing arts. With only that factor to rely on, the live performing arts, as Baumol and Bowen (1966, 165) emphasize, "cannot hope to match the remarkable record of productivity growth achieved by the economy as a whole." As a result, cost per unit of output in the live performing arts is fated to rise continuously relative to costs in the economy as a whole. That, in brief, is the unavoidable consequence of productivity lag.

On the other hand, industries in the "progressive" sector, in which productivity rises at a substantial rate, find themselves in a very favourable position. They can raise wages each year at the same rate at which productivity improves without increasing their unit labour costs at all. Hence, their prices need not rise even though their wages do.

A Numerical Example

In the accompanying table the upper panel shows the situation in a hypothetical manufacturing industry where productivity is increasing. Assume that widgets are the product. Output per work-hour is therefore measured by widgets produced per worker per hour. The first row shows that output per worker rises from 20 widgets in 2010 to 24 in 2015, an increase of 20 percent. Wages, shown in the second row, rise at the same rate as productivity, increasing from $10 per hour in 2010 to $12 an hour in 2015. Unit labour cost, equal to wages per work-hour divided by output per work-hour, is shown in the third row. In 2010, unit labour cost = $10/20 widgets, or 50 cents per widget. In 2015, unit labour cost is unchanged. Though wages have risen 20 percent, so has output per work-hour, leaving unit labour cost still at 50 cents per widget. Thus, wages in a progressive industry can rise as fast as productivity without causing any increase in costs (Table 1).

The lower panel of the table shows the situation in a hypothetical symphony orchestra, a live performing arts institution in which productivity is stagnant. We assume the following production conditions. The orchestra consists of 100 musicians. It plays five concerts per week in a hall that seats 1,600. Potential admissions (the "output" of the orchestra in productivity terms) is therefore 8,000 per week. The musicians work a 40-h week. Output per work-hour of the orchestra is therefore 8,000/40, or 200 admissions. Since there are 100 musicians, output per work-hour per musician is two admissions. This is shown in the first row of the lower panel and is unchanged from 2010 to 2015.

The second row of the lower panel shows that wages per hour for players in the orchestra rose from $20 in 2010 to $24 in 2015, an increase of 20 percent that matches the upward movement of wages in the general economy. Unit labour costs

Baumol's Cost Disease, Table 1 Hypothetical Illustration of Productivity Lag and Cost Impact

Industry	2010	2015	Percent change 2010–2015
Widget manufacturer			
Output in widgets per workhour	20	24	+20
Wage per hour	$10	$12	+20
Unit labour cost per widget			0
Symphony orchestra			
Output in admissions per workhour	2	2	0
Wage per hour	$20	$24	+20
Unit labour cost per admission	$10	$12	+20

Capacity of concert hall = 1,600; concerts per week = 5; potential admissions per week = 8,000; number of musicians = 100; musician work hours per week = 40; orchestra hours per week = 4,000; output per work hour: admissions per week + orchestra hours per week = 8,000 ÷ 4,000 = 2

for the orchestra are shown in the third row. In 2010 hourly wages were $20 and output per workhour was two admissions, yielding $10 per admission. By 2015 wages had increased to $24 an hour, while output per work-hour remained at 2, so that unit labour cost increased to $12 per admission. These hypothetical numbers show that in the live performing arts, unit labour costs can rise over time by the same rate at which productivity gains in the arts lag behind those in the general economy.

Historical Evidence on Performing Arts Costs

The historical record strongly supports the hypothesis that because of productivity lag, unit costs in the live performing arts increase substantially faster than the general price level does. Baumol and Bowen (1966) themselves tracked a great deal of this evidence. Their earliest cost data are for productions at the Drury Lane Theatre in London in the eighteenth century. They compared average cost per performance at the Drury Lane in the seasons 1771–1772 through 1775–1776 with costs per performance of the Royal Shakespeare Theatre in 1963–1964. In that period of almost two centuries, cost per performance multiplied 13.6 times. Over the same period a historical index of overall British prices shows them to have increased only 6.2 times. These increases can also be expressed as compound annual rates of growth, that is, as the annual growth rate that, if applied to the starting figure and compounded over the period in question, would result in the indicated final magnitude. On that measure, theatrical costs increased 1.4 percent per year while the annual rate of increase for the general price level was only 0.9 percent.

In the United States, Baumol and Bowen put together a nearly continuous cost history for the New York Philharmonic Orchestra beginning in 1843. Between that date and 1964, cost per concert rose at a compound annual rate of 2.5 percent, while the US index of wholesale prices rose an average of 1.0 percent per year. As Baumol and Bowen point out, the apparently small difference between these numbers leads to a startling divergence in costs when compounded decade after decade: The orchestra's cost per concert multiplied 20 times over in 121 years, while the general price level only quadrupled.[7]

For the years after World War II, Baumol and Bowen analyzed data on 23 major US orchestras, three opera companies, one dance company, and a sample of Broadway, regional, and summer theatres. In every group, the same results showed up: Cost per performance increased far more rapidly than the general price level. Moreover, they found a pattern in the postwar experience of Britain's Royal Shakespeare Theatre and London's Covent Garden (venue for the Royal Opera and Royal Ballet) so strikingly similar to US experience that they were encouraged to speculate that the structural problem of production in the live performing arts is one "that knows no national boundaries."

Consequences of the Cost Disease

The facts of productivity lag seem to be obvious. Evidence cited above indicates that it causes costs, and presumably prices, in the stagnant sectors to

rise relative to costs in the general economy, and that in the long run an extraordinary divergence in prices can occur. But why should we worry about it? After all, many service activities besides the arts are afflicted with productivity lag. It takes a barber just as long to cut hair, or a fine restaurant just as long to prepare and serve a gourmet meal, now as it did 50 years ago. Consequently, the prices of those services (and many others in which technological improvements are absent or unimportant) have risen far more rapidly than the general price level. Yet we hear no outcry about a haircutting crisis or an impending financial collapse of the gourmet restaurant industry. Why should we worry about productivity lag in the live performing arts, education, and health care? Why not let these sectors suffer whatever consequences the uneven progress of technological change metes out for them?

The answer must be that we as a society have a special interest in these sectors and that we are therefore unwilling to leave their fate to the dictates of the market as we do haircuts and gourmet meals. Implications for social capital and external costs and benefits are addressed in other entries in this dictionary. At this stage we simply explain the two principal points made by those who are concerned about the effects of productivity lag.

First, as we have already seen, productivity lag leads to steadily rising prices for the affected industries. This, in turn, makes it increasingly difficult to provide vital health care services to people of low or moderate income. In addition, it has been amply documented that those with low or even moderate incomes are grossly underrepresented in US arts audiences. Anyone who believes that this virtually automatic exclusion of the poor is socially undesirable is likely to be alarmed at the inexorable rise in ticket prices dictated by productivity lag.

The second unfortunate effect of productivity lag is that it puts the non-profit and government institutions responsible for most of our live performing arts, health care, and education under unremitting financial pressure (e.g., see Bates and Santerre 2015). Because relative costs are continuously increasing, they are under great pressure to raise ticket prices, health insurance rates, tuition fees, and taxes faster than the general rate of inflation, a strategy that is not easy to carry out, that many find philosophically repugnant, and that meets strong political opposition. While it is difficult to demonstrate rigorously, it seems reasonable to believe that a non-profit firm would find it easier to balance its budget in a technologically progressive industry, where unit costs are stable or falling year by year, than in a lagging one, where real costs are constantly moving upward and prices charged to customers must do likewise.

The financial problems facing performing arts groups as a result of productivity lag were emphasized by Baumol and Bowen. For them and for later writers, a company's "earnings gap," defined as the difference between its expenditures and its earned income, has appeared to be the most useful (though a far from unambiguous) measure of the financial strain it faces. In general the gap is covered by some combination of private donations and government subsidy.

Is There a "Quality Deficit"?

Faced with the continual upward pressure on costs generated by productivity lag, firms in the live performing arts might be expected to seek ways of economizing by gradually altering their choice of repertory or their production process. For example, theatrical producers might look for plays with smaller casts or plays that could be mounted with a single rather than multiple stage sets. Or they might try to compensate for higher costs by shunning artistically innovative plays that do not draw well at the box office and so have to be "carried" by revenues from more conventional offerings. We would expect this to occur most often in smaller cities where a single company might have a virtual monopoly on professional production. Orchestras and opera companies, too, might be driven away from innovative or "difficult" material by box office considerations. Or operating on the cost side, they might select programs with an eye to reducing rehearsal time or hire fewer outside soloists or other high-priced guest artists. Ballet companies could cut down on the use of specially commissioned music

or choreography and could eschew new productions that require elaborate sets or costumes.

Economic theory predicts and organizational behaviour demonstrates that firms will respond to rising input costs by economizing in their use of the offending inputs, but arts aficionados are likely to be disturbed when firms in the performing arts do just that. They are offended at the notion that *Hamlet* is no longer viable because its cast is too large, or that piano concertos will be less frequently heard because soloists have become too expensive (Rosen 1981). Hilda and William Baumol (1984) expressed their dismay at the notion that rising costs should narrow "the economically feasible range of artistic options." When that occurs it has been said that performing arts firms are reducing their fiscal deficit by incurring an "artistic deficit."

It is worth noting that, within the larger arts industry, this problem is most notable in the performing arts. In the visual arts – for example, in architecture – we fully expect practitioners to adapt their "products" to changes over time in the relative prices of alternative inputs. We are not surprised to find that modern buildings are devoid of the elaborate hand-carved stonework that decorated important buildings in earlier times. Indeed, the aesthetic rationale of the modern movement in architecture was precisely to design buildings that could use machine-finished materials in place of the increasingly costly hand-finished ones. In this instance it is not too strong to say that the necessity of adapting was the challenge that gave rise to a whole new school of design.

What makes the performing arts different is the fact that the past provides much of the substance that we wish to see performed. We do not want *Hamlet* with half the characters omitted because of the high cost of labour. Nor do we wish to give up symphony concerts in favour of chamber music recitals simply because symphonies employ too many musicians. We want the "range of artistic options" to include the option of hearing or seeing performances of great works that were invented under very different economic circumstances than our own. There would indeed be an artistic deficit if today's companies became financially unable to present for us the great works of the past.

Do performing arts institutions, responding to financial pressure, already exhibit an artistic deficit? Some of the available evidence is anecdotal, but there is also more systematic evidence. Hilda and William Baumol (1984) found that average cast size for all non-musicals produced on Broadway fell from 15.8 in 1946–1947 to 8.1 in 1977–1978. More recently, a study of opera repertory in the United States has shown that from 1983 to 1998 companies have increasingly produced popular operas at the expense of new or less well-known works. This could be interpreted as evidence of a growing artistic deficit in that field.

Although this section has focused on a quality deficit in the arts, it is easy to imagine the deleterious effects of a corresponding quality deficit in health care and education, where the current and future quality of life hangs in the balance.

Cures or Offsets to the Cost Disease?

As indicated earlier in this entry, a number of factors can work to offset the effects of productivity lag and the resulting cost impact, although it is not clear any of them can be a genuine and lasting cure. These are high income elasticity of demand for arts, education, and health care; economies of scale; rising revenue from related sources; and technology adoption in substitute products. Economies of scale were considered earlier; here we consider each of the other potentially offsetting influences.

Income Elasticity of Demand To the extent that the arts, higher education, and some health care can be regarded as "luxury" goods, spending in these areas rises faster than consumer incomes. A steadily growing economy, combined with a reasonably equitable distribution of buying power, should result in greater overall spending on these luxuries. Cost may go up, but consumers absorb the higher costs as they shift spending in the direction of luxury goods.

Revenue from Related Sources Earned income from related sources such as museum and theatre shops, sales of recordings, and special fund-raising

events, clearly do not enhance productivity, but to the extent they help to cover costs, ticket-buyers do not bear the full brunt of the cost disease.

Technology Adoption in Substitute Products Enhancements in music recording and distribution technologies can make orchestral concerts available in the living room or automobile at a fraction of the price of attending an orchestra concert (Baumol and Baumol 1985; Gordon 2016, 13). However, this is a cure or offset only to the extent that live attendance and audio or video recordings are actually substitutes, for which little evidence exists.

Conclusion

This entry introduced the problems associated with productivity lag in so-called stagnant sectors, with emphasis on the performing arts. Rising costs and higher ticket prices threaten to reduce the audience for the arts, but these difficulties may be at least partially offset in a growing economy by rising consumer incomes, an increasing taste for the arts, and falling unit costs attributable to economies of scale. One can thus remain guardedly optimistic about the continued financial viability of the performing arts. However, there is considerable evidence that growth of the arts earnings gap has been forestalled in part by an increasing artistic deficit.

Finally, we note implications for overall economic well-being: "In 2014, fully two-thirds of consumption spending went for services, including rent, health care, education, and personal care" (Gordon 2016, 578). And, we can add, the arts. This shift in spending toward the stagnant sectors implies a decline in future economic growth and diminished expectations of rising standards of living as conventionally measured.

Bibliography

Bates, L.J., and R.E. Santerre. 2015. Does Baumol's cost disease account for nonfederal public-sector cost growth in the United States? A new test of an old idea. *Social Science Quarterly* 96 (1): 251–258.

Baumol, W.J. 1967. Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review* 57 (3): 415–426.

Baumol, W.J. 1996. Children of performing arts, the economic dilemma: The climbing costs of health care and education. *Journal of Cultural Economics* 20 (3): 200–203.

Baumol, W.J. 2012. *The cost disease*. New Haven: Yale University Press.

Baumol, Hilda, and W.J. Baumol. 1984. The mass media and the cost disease. In *The economics of cultural industries*, ed. W.S. Hendon et al., 109–123. Akron: Association for Cultural Economics.

Baumol, Hilda, and W.J. Baumol. 1985. The future of the theater and the cost disease of the arts. In *Bach and the box: The impact of television on the performing arts*, ed. Mary Ann Hendon et al., 7–31. Akron: Association for Cultural Economics.

Baumol, W.J., and W. Bowen. 1966. *Performing arts: The economic dilemma*. New York: Twentieth Century Fund.

Bosworth, B.L., and J.E. Triplett. 2007. Services productivity in the United States: Griliches's services volume revisited, chapter 14. In *Hard-to-measure goods and services: essays in honor of Zvi Griliches*, ed. E.R. Berndt and C.R. Hulten. Chicago: University of Chicago Press.

Brooks, A.C. 1997. Toward a demand-side cure for cost disease in the performing arts. *Journal of Economic Issues* 31 (1): 197–207.

Cowen, Tyler. 1997. Why I do not believe in the cost-disease. *Journal of Cultural Economics* 20: 207–214.

Eldridge, L.P., and Jennifer Price. 2016. Measuring quarterly labor productivity by industry. Monthly Labor Review, 1–26 June. Retrieved at https://www.bls.gov/opub/mlr/2016/article/pdf/measuring-quarterly-labor-productivity-by-industry.pdf

Flanagan, Robert. 2012. *The perilous life of symphony orchestras*. New Haven: Yale University Press.

Gordon, R.J. 2016. *The rise and fall of American growth*. New Jersey: Princeton University Press.

Gray, C.M., K.J. Borowiecki, and James Heilbrun. Forthcoming. *The economics of art and culture*. 3rd ed. Cambridge University Press.

Hartwig, Jochen. 2008. Productivity growth in service industries: Are the transatlantic differences measurement-driven? *Review of Income and Wealth* 54 (3): 494–505.

Last, A.-K., and Heike Wetzel. 2011. Baumol's cost disease, efficiency, and productivity in the performing arts: An analysis of German public theaters. *Journal of Cultural Economics* 35: 185–201.

Rosen, Sherwin. 1981. The economics of superstars. *American Economic Review* 71 (5): 845–858.

Towse, Ruth, ed. 1997. *Baumol's cost disease, the arts and other victims*. Cheltenham: Edward Elgar.

Triplett, J.E., and B.P. Bosworth. 2003. Productivity measurement issues in service industries: "Baumol's disease" has been cured. *Economic Policy Review*, Federal Reserve Bank of New York.

Bayes, Thomas (1702–1761)

D. V. Lindley

Keywords

Bayes, T.; De Finetti, B.; Expectations; Inference; Laplace, P. S.; Price, R.; Probability; Subjective probability

JEL Classifications

B31

The Rev. Thomas Bayes was the eldest son of Joshua Bayes, a minister in the nonconformist church. He was probably educated at Coward's Academy. After assisting his father as pastor in Hatton Garden, London, he became, in 1731, Presbyterian minister at Mount Sion, Tunbridge Wells where he remained until his death on 17 April 1761. His fame today rests entirely on one paper, found by his friend Richard Price amongst Bayes' effects after his death and presented to the Royal Society (Bayes 1763; a convenient recent reference is Bayes 1958). The paper appears to have aroused little interest at the time and a proper appreciation was left to Laplace. Even today there is much discussion over just what Bayes meant, but the fact that so much interest is taken in a paper over 200 years old testifies to the importance of the problem and the brilliance of Bayes' argument.

The problem was this (as stated at the beginning of the paper): '*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.'

Bayes' solution depended on two original ideas. The first, in the modern notation where $p(A|B)$ means the probability of A given B, says

$$p(B|A) = p(A|B)p(B)|p(A)$$

and is always known as Bayes' theorem. The second idea is more controversial and open to many interpretations. The question is what 'rule is the proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it'?

To solve the problem Bayes took A to be the event of r happenings and s failures; B to be the unknown value θ of 'its happening in a single trial' so that $p(r, s|\theta) = \theta^r(1 - \theta)^s$; and supposed $p(r, s) = (r + s)^{-1}$ as a solution to the second question. This is equivalent to taking $p(\theta)$ as constant.

The importance of Bayes' ideas goes beyond the initial problem. Let A be any *particular* event and B some *general* proposition. Then his theorem enables one to pass from the probability of the particular given the general, $p(A|B)$, which, as above, is often straightforward, to the difficult probability of the general given the particular, $p(B|B)$. As such it provides a solution to the central problem of induction or inference, enabling us to pass from a particular experience to a general statement. This Bayesian inference applies generally in science, economics and law. A special case with statistical problems is called Bayesian Statistics. It has been shown by Ramsey (1931), De Finetti (1974/5) and others that this is the only coherent form of inference. Despite this, eminent philosophers like Popper (1959) still misunderstand Bayes and deny probabilistic induction.

Bayes' solution to the second question has not been generally accepted and the probability to be assigned to the general proposition before the particular is observed, $p(B)$, has been the subject of much discussion. Solutions by Jeffreys (1985), and by Jaynes (1983) using entropy ideas, have all met with difficulties. The best solution currently available is to accept that *all* probabilities are subjective so that, in particular, $p(B)$ is the subject's probability for the general proposition. This view is primarily due to De Finetti. Enough data (in the form of particular events) enable subjects, despite differences in $p(B)$, to have close agreement on $p(B|A)$.

An interesting feature of Bayes' approach is that he defines probability in terms of expectation. The amount you would pay for the expectation of one unit of currency were B to occur is $p(B)$.

Because of its confusion with utility concepts, this approach has not been much used.

It is hard to think of a single paper that contains such important, original ideas as does Bayes'. His theorem must stand with Einstein's $E = mc^2$ as one of the great, simple truths.

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* 53: 370–418.

Bayes, T. 1958. Reprint of the above with biographical note by G.A. Barnard. *Biometrika* 45: 293–315.

De Finetti, B. 1974/5. *Theory of probability,* 2 vols. New York: Wiley.

Jaynes, E.T. 1983. In *Papers on probability, statistics and statistical physics*, ed. R.D. Rosenkrantz. Dordrecht: Reidel.

Jeffreys, H. 1985. *Theory of probability.* Oxford: Clarendon Press.

Popper, K.R. 1959. *The logic of scientific discovery.* London: Hutchinson.

Ramsey, F.P. 1931. *The foundations of mathematics and other logical essays*. London: Kegan, Paul, Trench, Trubner.

Bayesian Econometrics

Dale J. Poirier

Abstract

'Bayesian econometrics' consists of the tools of Bayesian statistics applicable to economic phenomena. The Bayesian paradigm interprets 'probability' as a measure of 'uncertainty' or 'degree of belief' associated with the occurrence of a particular uncertain event, given the available information and any accepted assumptions. It prescribes how an individual *should* act in the face of such uncertainty in order to avoid undesirable inconsistencies. The coherence of the Bayesian approach contrasts sharply with conventional statistical methods which sometimes advocate negative estimators of positive quantities to ensure unbiasedness, and confidence intervals which may be null or consist of the whole parameter space.

Keywords

Bayes, T; Bayes' theorem; Bayesian econometrics; Bernoulli, J; Collinearity; de Finetti, B; Empirical Bayes analysis; Exchangeability; Expected subjective utility; Extreme bounds analysis; Frequentist statistics; Good, I.J; Hypothesis testing; Interval estimation; Jeffreys' rule; Laplace, P.S; Likelihood principle; Lindley, D; Markov chain Monte Carlo methods; Maximum likelihood; Model building; Objective probability; Point estimation; Prediction; Probability; Regression; Representation theorem; Savage, L. J; Statistical inference; Subjective probability; Uncertainty

JEL Classification

C11

'Bayesian econometrics' consists of the tools of Bayesian statistics applicable to economic phenomena. Bayesian statistics traces its roots back to Reverend Thomas Bayes (born circa 1702 and died in 1761) who was an ordained nonconformist minister in England. His ideas appear to have been independently developed by James Bernoulli, and later popularized independently by Pierre Laplace later in the 18th century. After more than a century of neglect, a rebirth of Bayesian statistics occurred in the 1930s at the hands of Sir Harold Jeffreys and Bruno de Finetti, and momentum built in the 1950s as a result of the efforts of I.J. Good, Dennis Lindley and Leonard J. Savage. Bayesian econometrics started in the 1960s with the work of Jacque Dreze and Arnold Zellner. With the computational revolution sparked by Markov chain Monte Carlo (MCMC) techniques in the 1980s and 1990s, many computational constraints were removed, and Bayesian analysis was flourishing in a wide variety disciplines as the new millennium began.

The Bayesian paradigm interprets 'probability' as a measure of 'uncertainty' or 'degree of belief' associated with the occurrence of a particular uncertain event, given the available information

and any accepted assumptions. It prescribes how an individual *should* act in the face of such uncertainty in order to avoid undesirable inconsistencies.

Consider an individual asked to quote probabilities on a set of uncertain events, and required to accept any wagers about these events. According to Bruno de Finetti's *coherency principle*, such an individual should never assign probabilities so that someone else can select stakes that guarantee a sure loss (*Dutch book*) for the individual whatever the eventual outcome. This simple principle implies the usual axioms of probability except that the additivity of probability for unions of disjoint events is required to hold only for finite unions.

Expected utility maximization (or loss minimization) provides a basis for rational decision making, and Bayes' theorem describes how beliefs evolve as data are obtained. There are numerous axiomatic formulations leading to the central unifying Bayesian prescription of maximizing expected subjective utility as the guiding principle of Bayesian statistical analysis. Bernardo and Smith (1994, ch. 2) is a valuable introduction to this vast literature. While the *descriptive* accuracy of the Bayesian approach in capturing the actual behaviours of individuals is questioned by many opponents, Bayesians claim that the Bayesian view provides only *normative* guidelines for behaviour.

The *subjective* interpretation of probability is based on an individual's personal assessment of a situation. For evidence of the use of subjectivity by history's most illustrious scientists, see Press and Tanur (2001). Accordingly, probability is a property of an individual's perception of reality. In contrast, according to objective interpretations, probability is a property of reality itself. For subjectivists there are no 'true unknown probabilities' in the world to be discovered. Instead, 'probability' is in the eye of the beholder. In de Finetti's words, 'Probability does not exist'.

De Finetti assigned a fundamental role in Bayesian analysis to exchangeability. A finite sequence of random quantities is *exchangeable* if the joint probability of the sequence, or any subsequence, is invariant under permutations of the subscripts. An infinite sequence is exchangeable if any finite subsequence is exchangeable. Exchangeability involves recognizing symmetry in beliefs concerning *observables*, and presumably this is something about which a researcher may have intuition. It provides an operational meaning to the weakest possible notion of a sequence of 'similar' random quantities. It is operational because it requires only probability assignments of observable quantities, although admittedly this becomes problematic in the case of infinite exchangeability.

The links between exchangeable beliefs over uncertain *observables* and the parameters in statistical models are provided by various generalizations of Bruno de Finetti's celebrated representation theorem for infinite sequences of exchangeable Bernoulli random variables (see Bernardo and Smith 1994, ch. 4). These theorems provide conditions under which exchangeability, and other symmetries, give rise to an isomorphic world consisting of i.i.d. observations with a given sampling distribution, conditional on a mathematical construct (a parameter), and guarantee the existence of a prior distribution for it. De Finetti put parameters in their proper perspective: they are mathematical constructs that provide a convenient index for a family of probability distributions, and they induce conditional independence in sequences of observables.

Bayesian inference involves updating prior beliefs into posterior beliefs conditional on observed data. Appealingly, Bayesian analysis requires only a few general principles that are applied over and over again in different settings. Bayesians begin by specifying a joint distribution for all quantities (denoted in bold italics) under consideration except known constants. The Bayesian paradigm reduces statistical inference to applied probability. Quantities that become known under sampling (data) are denoted by the T-dimensional vector $y \in Y$, and the remaining unknown (and unobserved) quantities (parameters) by the m-dimensional vector $\theta \in \Theta \subseteq R^m$. Unless noted otherwise, y and θ are treated as continuous random variables. Working in terms of densities, consider

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\theta}) &= f(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \\ &= f(\boldsymbol{\theta}|\mathbf{y}) f(\mathbf{y}), \mathbf{y}, \boldsymbol{\theta} \in \mathbf{Y} \times \boldsymbol{\Theta}, \end{aligned} \quad (1)$$

where f(θ) is the *prior density*, f(**y**|**θ**) viewed as a function of θ for known y is the *likelihood* function [denoted L(**θ**, *y*)], f(**θ**|**y**) is the *posterior density*, and

$$f(\mathrm{y}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) d\boldsymbol{\theta} = \mathbf{E}_{\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{\theta}, y)], \mathbf{y} \in \mathbf{Y}, \quad (2)$$

is the *marginal density* of the data y. From (1), Bayes' theorem for densities follows:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}, \mathbf{y})}{f(\boldsymbol{\theta})} \sim \propto f(\boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}, y), \boldsymbol{\theta} \in \Theta. \quad (3)$$

Hereafter, (3) is adopted as the way to update prior beliefs when $y = \mathrm{y}$ is observed.

Fortunately, sometimes the integration in (2) can be performed analytically and so the updating of prior beliefs in light of the data to obtain the posterior beliefs is straightforward. These situations correspond mostly to cases where L(θ, *y*) belongs to the exponential family of densities. In this case the prior density can be chosen so that the posterior density falls within the same elementary family of distributions as the prior. These prior families are called *conjugate families*. Conjugate priors are more flexible than they may appear at first since mixtures of conjugate priors are themselves conjugate, although they may be daunting to elicit.

The denominator in (3) serves as an integrating constant. Hence, when one considers experiments employing the same prior, and which yield proportional likelihoods for the observed data, identical posteriors will emerge, consistent with the *likelihood principle* (Berger and Wolpert 1988). Unlike the inherent *ex ante* perspective of frequentist statistics, which seeks properties of procedures in repeated sampling, posterior density (3) is *ex post* – it conditions on the observed data $y = \mathrm{y}$, and dispenses with the part of the sample space Y that could have been observed but was not.

In most practical situations not all elements of θ are of direct interest. Let $\boldsymbol{\theta} = [\beta', \delta']' \in \mathbf{B} \times \boldsymbol{\Delta}$ be partitioned into *parameters of interest* β and *nuisance parameters* δ. Nuisance parameters are well-named for frequentists, because dealing with them in a general setting is one of the major problems non-Bayesian researchers face. In contrast, Bayesians adopt a universal approach to eliminating nuisance parameters from the problem: integrate them out of the joint posterior to obtain the marginal posterior density for *β*:

$$f(\beta|\mathbf{y}) = \int_{\boldsymbol{\Delta}} f(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{y}) d\boldsymbol{\delta}, \boldsymbol{\beta} \in \mathbf{B}. \quad (4)$$

Point Estimation

Consider a *loss* (*cost*) function $C(\widehat{\beta}, \beta)$ for the parameters of interest *β*, that is, a nonnegative function satisfying $C(b, b) = 0$ and which measures the consequences of using the estimate $\widehat{\beta}$ when the parameter of interest is *β*. Both frequentists and Bayesians seek to 'minimize' (in some sense) $\mathbf{C}(\widehat{\beta}, b)$, but first its randomness must be eliminated.

From the frequentist point of view, *β* is a degenerate random variable equal to *β*, but $\mathbf{C}(\widehat{\beta}, b)$ is stochastic because $\widehat{\beta}$ is viewed *ex ante* as the estimator $\widehat{\beta} = \widehat{\beta}(y)$ depending on the data *y* which are random viewed *ex ante*. One way to circumscribe the randomness of $C(\widehat{\beta}, b)$ is to focus on its expected value, assuming it exists. Frequentists consider the risk function

$$\mathbf{R}\left(\widehat{\beta}|\beta, \boldsymbol{\delta}\right) = \mathrm{E}_{y|\beta=\beta, \boldsymbol{\delta}=\boldsymbol{\delta}}\left[C\left(\widehat{\beta}(\mathbf{y}), \beta\right)\right], \quad (5)$$

where the expectation is taken with respect to the sampling density f(y|*β*, δ), $\mathrm{y} \in \mathrm{Y}$.

In contrast, the Bayesian perspective is entirely *ex post*, and it seeks a function $\widehat{\beta} = \widehat{\beta}(\mathrm{y})$ of the observed data $y = \mathrm{y}$ to serve as a point estimate of the parameter of interest *β*. Unlike the frequentist approach, no role is provided for data that could have been observed, but were not. Since *β* is unknown, the Bayesian perspective suggests formulation of subjective beliefs about it, given all the information at hand. Such information is fully contained in

marginal posterior density (4). In contrast to (5), Bayesians focus on *expected posterior loss*:

$$\mathbf{c}\left(\widehat{\beta}|\mathbf{y}\right) = \mathrm{E}_{b|y=y}\left[\mathbf{C}\left(\widehat{\beta}, b\right)\right] = \int_{\mathbf{B}} \mathrm{C}\left(\widehat{\beta}, \beta\right) \mathrm{f}(\beta|\mathbf{y})\mathrm{d}\beta. \qquad (6)$$

The second Bayesian commandment (after Bayes' theorem) is: act so as to minimize expected posterior loss, that is, find $\widehat{\beta}_* \equiv \mathbf{argmin}_{\widehat{\beta}} \mathrm{E}_{b|\mathbf{y}}[\mathrm{C}(\widehat{\beta}, b)]$. Frequentists emphasize the sampling distribution $\boldsymbol{y}|\beta = \beta, \boldsymbol{\delta} = \delta$ and Bayesians emphasize the posterior distribution $\boldsymbol{\beta}|\boldsymbol{y} = \mathrm{y}$. The debate is about the desired conditioning – as are most debates in statistics. Posterior expectation (8) removes β from $\mathbf{C}(\widehat{\beta}, b)$ yielding a criterion $\mathbf{c}\ (\widehat{\beta}|\mathrm{y})$, unlike risk function (5), involving only known quantities.

For simplicity, consider univariate β and the following three loss functions in which c, c_1, c_2, and d are known constants: the *quadratic loss function* $\mathbf{C}\ (\widehat{\beta}, b) = (\widehat{\beta} - b)^2$, the *asymmetric linear loss function* $\mathbf{C}(\widehat{\beta}, b) = c_1|\widehat{\beta} - b|$, if $\widehat{\beta} \leq b$, and $\mathrm{C}(\widehat{\beta}, b) = c_2|\widehat{\beta} - b|$ if $\widehat{\beta} > b$, and the *all-or-nothing loss function* $\mathbf{C}(\widehat{\beta}, b) = \mathrm{c}$, if $|\widehat{\beta} - b| > \boldsymbol{d}$, and $\mathrm{C}(\widehat{\beta}, b) = 0$, if $|\widehat{\beta} - b| \leq \boldsymbol{d}$. The resulting Bayesian point estimates are the posterior mean, the qth posterior quantile where $q = \frac{c_1}{c_1 + c_2}$, and the centre of an interval of width 2d having maximum posterior probability (yielding the posterior mode as $\mathrm{d} \to 0$ respectively. When β is a vector, the most popular loss functions are the *weighted squared error* generalization of quadratic loss, $\mathbf{C}(\widehat{\beta}, b) = (\widehat{\beta} - b)'\mathrm{Q}(\widehat{\beta} - b)$, where Q is a positive definite matrix, or the all-or-nothing loss function. In these cases the Bayesian point estimates are again the posterior mean and mode (as $\mathrm{d} \to 0$), respectively.

Minimum risk estimators do not exist in general because (5) depends on β and δ, and so an estimator that minimizes (5) will also depend on β and δ. Often extraneous side conditions are imposed (for example, unbiasedness) to sidestep the problem. In contrast, Bayesian point estimates are optimal by construction from the *ex post* standpoint. In general they also have good *ex ante* risk properties. Consider the minimizer of (6) viewed from the *ex ante* standpoint before the data are realized, that is, the *Bayesian point estimator* $\widehat{\beta}_* = \widehat{\beta}_*(y)$. Provided the prior distribution is *proper* (it integrates to unity), then $\widehat{\beta}_*(y)$ satisfies the minimal frequentist requirement of *admissibility* (its risk cannot be dominated by another estimator everywhere in the parameter space). Furthermore, in most interesting settings, all admissible estimators are either Bayes or limits thereof known as *generalized Bayes estimators* based on an *improper prior* whose integral diverges.

Interval Estimation

Bayesian interval estimation follows directly from the posterior density $\mathrm{f}(\beta|\mathrm{y})$. Because opinions about the unknown parameter are treated in a probabilistic manner, there is no need to introduce the additional concept of 'confidence'. For example, given a region $\mathbf{B}^\dagger \subset \mathrm{B}$, it is meaningful to ask: given the data, what is the *probability* that β lies in $\mathbf{B}^\dagger$? The answer is direct:

$$\mathbf{Prob}\left(b \in \mathbf{B}^\dagger|\mathbf{y}\right) = \int_{\mathbf{B}^\dagger} \mathrm{f}(\boldsymbol{\beta}|\mathbf{y})\mathrm{d}\beta. \qquad (7)$$

Alternatively, given a desired probability content of $1 - \alpha$, it is possible to reverse this procedure and find a corresponding region $\mathbf{B}^\dagger$. The 'smallest' region $\mathbf{B}^\dagger$ satisfying (9), known as the *highest posterior density* (*HPD*) *region of content* $(1 - \alpha)$ *for* β corresponds to imposing the added condition that for all $\beta_1 \in \mathbf{B}^\dagger$ and $\beta_2 \notin \mathbf{B}^\dagger$, $\mathrm{f}(\beta_1|\mathrm{y}) \geq \geq \mathbf{f}\,(\beta_2|\mathrm{y})$.

Hypothesis Testing

Consider a partition of the parameter space B for the parameter of interest β according to $\mathrm{B} = \mathrm{B}_1 \cup \mathrm{B}_2$, where $\mathrm{B}_1 \cap \mathrm{B}_2$ is null. Suppose interest lies in testing H_1: $\beta \in \mathrm{B}_1$ versus H_2: $\beta \in \mathrm{B}_2$ based on a sample y yielding the likelihood $\mathrm{L}(\beta, \delta; \mathrm{y})$. The relevant decision space is $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2\}$, where $\mathrm{d}_\mathrm{j} \equiv$ choose hypothesis H_j $(\mathrm{j} = 1, 2)$. Extensions to cases involving more than

two hypotheses are straightforward. Let C(d, b) ≥ 0 denote the relevant loss function. Without loss of generality, assume that correct decisions yield zero loss.

From the Bayesian perspective a hypothesis is of interest only if the prior distribution assigns it positive probability. Therefore, assume $\underline{\pi}_j \equiv \mathbf{Prob(H_j)} = \mathbf{Prob}(b \in \mathbf{B_j}) > \mathbf{0}$ $\mathbf{(j = 1,2)}$ with $\underline{\pi}_1 + \underline{\pi}_2 = 1$. Let $f_j(\beta, \delta|H_j)$ be the prior density under H_j (j = 1, 2). Under H_j, the marginal data density (*expected likelihood*) is

$$f(\mathbf{y}|\mathbf{H}_j) = \int_{\Delta}\int_{\mathbf{B}_j} \mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\delta}; \mathbf{y}) dF_j(\beta, \delta|\mathbf{H}_j) = E_{b,d|H_j}[\mathscr{L}(b, d; \mathbf{y})](\mathbf{j} = 1, 2), \quad (8)$$

where $F_j(\cdot)$ denotes the c.d.f. corresponding to the distribution β, $\delta \mid H_j$. From Bayes' theorem it follows that the posterior probability of H_j is

$$\frac{\overline{\pi}_j = \text{Prob}(H_j|y) = \underline{\pi}_j f(y|H_j)}{f(y)(j = 1, 2),} \quad (9)$$

where the marginal density of the data is $f(y) = \underline{\pi}_1 f(y|H_1) + \underline{\pi}_2 f(y|H_2)$. Under H_j, the posterior density of β and δ is (according to Bayes' theorem):

$$f(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{y}, \mathbf{H}_j) = \frac{f_j(\beta, \delta|H_j)\mathcal{L}(\beta, \delta; y)}{f(y|H_j)}, \beta \in \mathbf{B}_j, \boldsymbol{\delta} \in \boldsymbol{\Delta}(j = 1, 2). \quad (10)$$

As in the case of estimation, the optimal Bayesian decision $\mathbf{d}_*$ in the hypothesis testing context minimizes expected posterior loss, that is, $\mathbf{d}_* \equiv \text{argmin}_d\, c(d|\mathbf{y})$, where

$$c(d|\mathbf{y}) = \overline{\pi}_1 c(d|\mathbf{y}, H_1) + \overline{\pi}_2 c(d|\mathbf{y}, H_2), \quad (11)$$

and $c(d|y,H_j) = E_{\theta|y,Hj}[C(d;\theta](j = 1,2)$. Specifically, $c(d_1|\mathbf{y}) = \overline{\pi}_2 c(d_1|\mathbf{y}, H_2)$, and $c(d_2|\mathbf{y}) = \overline{\pi}_1 c(d_2|\mathbf{y}, H_1)$. Therefore, it is optimal to choose $\mathbf{H}_2$[that is, $c(d_2|y) < c(d_1|y)$] iff

$$\mathbf{d}_* = d_2 \text{ iff } \frac{\overline{\pi}_2}{\overline{\pi}_1} > \frac{c(d_2|\mathbf{y}, H_1)}{c(d_1|\mathbf{y}, H_2)} \quad (12a)$$

The quantities $\frac{\overline{\pi}_2}{\underline{\pi}_1}$ and $\frac{\overline{\pi}_2}{\overline{\pi}_1}$ are the the *prior odds and posterior odds*, respectively, of H_2, versus H_1 from (9) it follows immediately that these two odds are related by $\frac{\overline{\pi}_2}{\overline{\pi}_1} = B_{21}\left(\frac{\overline{\pi}_2}{\underline{\pi}_1}\right)$, where $B_{21} = \frac{f(y|H_2)}{f(y|H_1)}$ is the *Bayes factor for H_2 versus H_1*. See Kass and Raftery (1995) for an excellent review. In terms of the Bayes factor B_{21}, (12a) can also be written

$$\mathbf{d}_* = d_2 \text{ iff } B_{21} \geq \left[\frac{c(d_2|\mathbf{y}, H_1)}{c(d_1|\mathbf{y}, H_2)}\right]\left[\frac{\underline{\pi}_2}{\underline{\pi}_1}\right]. \quad (12b)$$

In general, expected posterior loss c(d|y, H_j) depends on the data y, and hence, Bayes factor B_{21} does *not* serve as complete data summary because the right-hand side of the inequality in (12b) also depends on the data. One exception is when both hypotheses are simple. Another is when an all-or-nothing loss is used such that the loss $\mathbf{C}(d_i, b) = \overline{\mathbf{C}}_i$ resulting from decision d_i when $\beta \in B_j$, $i \neq j$, is constant for all $\beta \in B_j$. In this case, for $i \neq j$, $c(d_1|y, H_j) = c(d_i|y, H_j) = C_i$, and decision rule (12b) reduces to

$$\frac{\mathbf{d}_* = d_2 \text{ iff } B_{21} \geq \underline{\pi}_1 \overline{\mathbf{C}}_2}{\underline{\pi}_2 \overline{\mathbf{C}}_1.} \quad (12c)$$

The right-hand side of the inequality in (12c) is a known constant *Bayesian critical value*.

Prediction

The sampling distribution of an out-of-sample $\tilde{y} \in \tilde{\mathbf{Y}}$ given $y = y$ and θ, would be an acceptable predictive distribution if θ was known, but without knowledge of θ it cannot be used. In its place is the *Bayesian predictive density*

$$\begin{aligned} f(\tilde{\mathbf{y}}|\mathbf{y}) &= \frac{f(\tilde{\mathbf{y}}, \mathbf{y})}{f(\mathbf{y})} = \int_{\boldsymbol{\Theta}} \frac{f(\tilde{\mathbf{y}}, \mathbf{y}, \boldsymbol{\theta})}{f(\mathbf{y})} d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} f(\tilde{\mathbf{y}}, \mathbf{y}, \boldsymbol{\theta}) \left[\frac{f(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y})}\right] d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} f(\tilde{\mathbf{y}}, \mathbf{y}, \theta) f(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \boldsymbol{E}_{q|\mathbf{y}}[f(\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\theta})], = \tilde{\mathbf{y}} \in \tilde{\mathbf{Y}} \end{aligned} \quad (13)$$

If the past and future are independent conditional on θ (as in random sampling), then f ($\tilde{\mathbf{y}}|\mathbf{y}$, $\boldsymbol{\theta}$) = f ($\tilde{\mathbf{y}}|\boldsymbol{\theta}$) Letting $\overline{\mathbf{C}}(\tilde{y}_p,\tilde{y})$ denote a *predictive loss function* measuring the performance of a predictor $\tilde{y}_p$ of $\tilde{y}$, the optimal point predictor $\tilde{\mathbf{y}}_*$ is defined to be $\tilde{\mathbf{y}}_* \equiv \text{argmin}_{\tilde{\mathbf{y}}_p} \boldsymbol{E}_{\tilde{\mathbf{y}}|\mathbf{y}}\left[\overline{\mathbf{C}}\left(\tilde{\mathbf{y}}_\mathbf{p},\mathbf{y}\right)\right]$. For example, if $\tilde{y}$ is a scalar and predictive loss is quadratic, then the optimal point estimate is the predictive mean $\tilde{\mathbf{y}}_* = \boldsymbol{E}(\tilde{\mathbf{y}}|\mathbf{y})$. Predictive density (13) can also be used to generate forecast intervals analogous to HPD intervals.

Predictive density (13) treats all parameters as nuisance parameters and integrates them out of the predictive problem. A similar strategy is used when adding parametric hypotheses to the analysis. Consider the hypothesis H_j and associated prior $f_j(\beta, \delta|H_j)$ (j = 1, 2). Given data y leading to the posterior $f_j(\beta, \delta|y, H_j)$, the predictive density of ~ y conditional on H_j is

$$f\left(\tilde{\mathbf{y}}|\mathbf{y},H_j\right) = \int_\Theta f\left(\tilde{\mathbf{y}}|\boldsymbol{\theta},\mathbf{y},H_j\right) f\left(\boldsymbol{\theta}|\mathbf{y},H_j\right) d\boldsymbol{\theta}\ \tilde{\mathbf{y}} \in \tilde{\mathbf{Y}}. \tag{14}$$

Using the posterior probabilities (9), the marginal predictive density of $\tilde{y}$ is the mixture density

$$f(\tilde{\mathbf{y}}|\mathbf{y}) = \overline{\pi}_1 f(\tilde{\mathbf{y}}|\mathbf{y},H_1) + \overline{\pi}_2 f(\tilde{\mathbf{y}}|\mathbf{y},H_2),\ \tilde{\mathbf{y}} \in \tilde{\mathbf{Y}} \tag{15}$$

and it is the basis for interval and point prediction. For example, under quadratic loss the optimal Bayesian point prediction is the predictive mean

$$E(\tilde{\mathbf{y}}|\mathbf{y}) = \overline{\pi}_1 E(\overline{\pi},H_1) + \overline{\pi}_2 E(\tilde{\mathbf{y}}|\mathbf{y},H_2), \tag{16}$$

which is a weighted average of the optimal point forecasts $E(\tilde{y}|y, H_1)$ under each hypothesis. The weights $\overline{\pi}_j$(j = 1, 2) in (16) have an intuitive appeal: the forecast of the more probable hypothesis a posteriori receives more weight.

Choice of Prior

Critics of Bayesianism find the choice of prior is the major stumbling block in adopting the Bayesian approach. In contrast, proponents see the required effort to be manageable and well worth it. Usually the likelihood is parameterized to facilitate thinking in terms of θ, and so subject matter considerations should suggest 'plausible' values of θ. Even when such direct thinking about θ is possible, it is also useful to think predictively (for example, see Kadane and Wolfson 1998) about the observable *y* and use (2) to back out a parametric prior f(θ|λ) for a specific value of some hyperparameter $\boldsymbol{\lambda} \in \Lambda$ in some space Λ. Usually such analyses restrict attention to conjugate priors. This ideal, however, is difficult to achieve.

Public research involving only a single prior is likely to draw few readers. Entertaining various professional positions in terms of θ can lead to different choices of **λ**. Rather than thinking of eliciting *the* prior, it is more useful to think in terms of a family F = {f(θ|λ), $\lambda \in \Lambda$} of parametric priors. If a prior f(λ) is available for **λ**, then we are back in the single prior case with *the* prior $f(\theta) = \int_\Lambda f(\theta)\, f(\lambda) d\lambda$. In most practical problems, however, there will be no agreed upon f(λ), and the researcher is left with investigating the sensitivity of the analyses to different elements in F. This is easier said than done, but in principle it can be done. For large dimensional θ, this can be difficult because the effects of the prior can be subtle: it may have little posterior influence on some functions of the data and have an overwhelming influence on other functions. Often a quantity of interest like the posterior mean E(θ|y) can be analytically restricted to a fairly small set of possible values for any given $\lambda \in \Lambda$. The *extreme bounds analysis* developed by Leamer (1982) is a leading example. In contrast, *empirical Bayes analysis* proceeds by using the data to estimate **λ**.

Kass and Wasserman (1996) survey formal rules that have been suggested for choosing a prior. Many of these rules reflect the desire to let the 'data speak for themselves'. This has led to variety of non-subjective priors intended to capture the elusive notion of *non-informativeness*. These priors are intended to lead to proper posteriors dominated by the data. They also serve as benchmarks for posteriors derived from ideal subjective considerations. At first many of these

priors were also motivated on simplicity grounds. But as problems were discovered, and other features were seen to be relevant, derivation of such priors became more complicated, possibly even more so than a legitimate attempt to elicit an actual subjective prior.

One interpretation of letting the data speak for themselves is to use classical techniques. Maximum likelihood estimates are rationalizable in a Bayesian framework by appropriate choice of prior distribution and loss function, specifically a uniform prior and an all-or-nothing loss function. But in what parameterization should one be uniform?

In order to overcome the re-parameterization problem, Jeffreys sought a general rule for choosing a prior so that the same posterior inferences were obtained regardless of the parameterization chosen. Jeffreys (1961) made a general (but not dogmatic) argument in favor of choosing a prior proportional to the square root of the information matrix, that is, $f(\theta) \propto |\mathbf{J}(\theta)|^{1/2}$, where $\mathbf{J}(\theta) \equiv E_{y|\theta}[-\partial^2 L(\theta; \boldsymbol{y})/\partial\theta\partial\theta']$ is the information matrix of the sample. This prior has the desirable feature that if the model is reparameterized by a one-to-one transformation, say $\psi = h(\theta)$, then choosing the prior $f(\psi) \propto |E_{y|\infty}[-\partial^2 L(\psi;\mathbf{y})/\partial\psi\partial\psi']|^{1/2}$ will lead to identical posterior inferences as using $f(\theta)$. Such priors are said to follow *Jeffreys' rule*.

Not all of Jeffreys' recommendations always followed Jeffreys rule: When Θ is finite, Jeffreys assigned equal probabilities to each of the values. When Θ is a bounded interval, Jeffreys assumed a constant proper prior. When $\Theta = R$, Jeffreys assumed a constant improper prior. When $\Theta = [0,\infty)$, Jeffreys chose $f(\theta) = \theta^{-1}$ because it is invariant under power transformations. When $\theta = [\theta_1, \theta_2]'$ where θ_1 is a location parameter and θ_2 is a non-location parameter, Jeffreys chose $\mathbf{f(\theta)} \propto |\mathbf{J(\theta)}j^{1/2}$, where $J(\theta)$ is calculated holding θ_1 fixed. In the case of mixture models, Jeffreys argued that the mixing parameters should be treated independently from the other parameters. There is a fair amount of agreement that such priors may be reasonable in one-parameter problems, but substantially less agreement (including Jeffreys) in multiple parameter problems.

Usually, Jeffreys' rule and other formal rules surveyed by Kass and Wasserman (1996), lead to *improper priors*, that is, priors which integrate to infinity rather than unity (a *proper prior*). When blindly plugged into Bayes' theorem as a prior they lead to proper posterior densities, but not always. They also produce proper predictive densities (13), but *not* proper marginal data densities (8). Furthermore, improper priors, in contrast to proper priors, are not guaranteed to lead to admissible Bayesian point estimators, and marginalization paradoxes can occur.

Bernardo (1979) suggested a method for constructing *reference priors* offering two innovations. First, he defined a notion of missing information in terms of the *Kullbach-Leibler distance* between the posterior and the prior density. Second, he developed a stepwise procedure for handling *nuisance parameters*. If there are no nuisance parameters, then his method usually leads to Jeffreys' rule. Subsequently, numerous refinements have been made in joint work with James O. Berger.

There are many candidates for non-subjective priors, and they often have properties that seem rather non-Bayesian. Most non-subjective priors depend on some or all of the following: (*a*) the form of the likelihood, (*b*) the sample size, (*c*) an expectation with respect to the sampling distribution, (*d*) the parameters of interest, and (*e*) whether the researcher is engaging in estimation, testing or predicting. The dependency in (*c*) of Jeffreys' prior on a sampling theory expectation makes it sensitive to a host of problems related to the likelihood principle. In light of (*d*), a *non-subjective* prior can depend on *subjective* choices such as which are the parameters of interest and which are nuisance parameters. Different quantities of interest require different non-subjective priors which cannot be combined in a coherent manner. My advice is use a non-subjective prior only with great care, and never alone. I include non-subjective priors in the class of priors over which I perform a sensitivity analysis.

One reaction to choice of prior is to not make one and proceed with an asymptotic analysis. The same way sampling distributions of the maximum

likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ in regular situations is asymptotically normal, posterior density (5) can be approximated as $T \to \infty$ by the multivariate normal density $\boldsymbol{\varphi}_{\mathbf{m}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{\mathbf{ML}}, [\mathbf{J}_{\mathbf{T}}(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})^{-1}]) = (2\pi)^{m/2} |\mathbf{J}_{\mathbf{T}}(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})|^{1/2} \exp[-\frac{1}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})' J_T(\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{\mathbf{ML}})]$, where $\mathbf{J}_{\mathbf{T}}(\cdot)$is the information matrix. This approximation does not depend on the prior. As an approximation to the posterior density of θ, the approximation usually improves by replacing the information matrix by the observed Hessian of the log-likelihood evaluated at $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$. The quality of this approximation can usually be improved by incorporating some information on the prior. For example, by using $\boldsymbol{\varphi}_{\mathbf{m}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}, [\overline{\mathbf{H}}_{\mathbf{T}}(\widehat{\boldsymbol{\theta}})]^{-1})$, where $\widehat{\theta}$ is the posterior mode and $\overline{H}_T(\widehat{\boldsymbol{\theta}})$ is the Hessian of the log posterior evaluated at $\widehat{\theta}$. Further asymptotic analysis using *Laplace approximations* (see Tierney and Kadane 1986) often given remarkably accurate results.

Model Building

A 'true model' is an oxymoron. An economic model is an abstract representation of reality that highlights what a researcher deems relevant to a particular economic issue. By definition an economic model is literally false, and so questions regarding its literal truth are trivial. Whether the model is useful is another matter.

A subjectivist's *econometric model* expresses probabilistically the researcher's beliefs concerning future observables of interest to economists. It has two components: a likelihood for viewing observables in the world, and a prior reflecting a professional position of interest. Poirier (1988) introduced the metaphor *window* for a likelihood function because it captures its essential role in de Finetti representation theorems: a parametric medium for viewing the observable world. Both model components are subjective, and both involve mathematical constructs called *parameters*. Parameters simply index distributions; any correspondence to physical reality is a rare side bonus.

In choosing the window L(θ, y) the researcher is torn in two directions: choosing the dimensionality of $\boldsymbol{\theta}$ to be large increases the chances of getting a bevy of researchers *to agree to disagree* in terms of the appropriate priors for θ, but a large dimensional $\boldsymbol{\theta}$ necessitate increasingly more informative priors if anything useful is to be learned from a finite sample. In one sense this dichotomy between prior and likelihood is tautological: if there is no agreement, then presumably the likelihood can always be expanded until agreement is obtained. The resulting window, however, may be hopelessly complex. The 'bite' in the statement comes from the assertion that a researcher believes agreement is compelling in the case of a particular window. Despite the many arguments in the literature over the wisdom of 'general to specific' as opposed to 'specific to general modelling', observed behaviour suggests researchers start with a finite parameterization of the problem that can be both simplified and expanded. The arguments are really over a matter of emphasis rather than kind.

Diagnostic checking of the maintained initial window can help achieve agreement on it. If the diagnostic checks indicate window expansion, then rethinking is required, a new window must be introduced, and the diagnostic checking process repeated. The extent of diagnostic testing depends in part on the size of the initial window. Everything else being equal, small windows require more checking to convince others of their value than large windows. Reporting that the initial window passes diagnostic checks is intended to soothe the concerns of members of the research community. For good discussions of diagnostic checking, see Gelman et al. (2003) and Lancaster (2004). Such checking can be as much an art as a science.

Conscientious empirical researchers provide their readers with a variety of ways of looking at the data. This amounts to checking how the observed data fit marginal density (2), how out-of-sample observables fit predictive densities (13) or (15), and how posterior densities (3) or (10) are summarized and interpreted. This task is complicated when m is large or when many hypotheses are entertained. Furthermore, the question arises: 'How should we bring together the results?' Is one hypothesis is to be chosen after an 'enlightened' search of the data? If so, then the

question is how to properly express uncertainty that reflects both sampling uncertainty from estimating the unknown parameters under a hypothesis and uncertainty over the hypothesis itself. The common practice of choosing a single hypothesis and then proceeding conditionally on it, is difficult to rationalize because the researcher's uncertainty is understated unless that hypothesis has a posterior probability near unity. Readers are interested in a clear articulation of the researcher's uncertainty because it can serve as a useful gauge or reference point for their own uncertainty.

When considering two hypotheses H_1 and H_2 it is possible to assign only $\underline{\pi}_1 + \underline{\pi}_2 = -\varepsilon$ prior probability to them, and to reserve ε $(0 < \varepsilon < 1)$ probability for an unspecified H_3 representing 'something else'. Then interpreting $\underline{\pi}_j$ relatively as Prob(H_j|H_1 or H_2) (j = 1, 2), posterior probabilities (11) can be computed and also interpreted relatively as Prob(H_j|y, H_1 or H_2) *without specifying* ε. If in the process the researcher's creative mind has a new insight leading to specification of 'something else', then some fraction $\underline{\pi}_3$ of $1-\varepsilon$ can be allocated to H_3 and the process repeated with the remaining portion allocated to a another unspecified H_4. The catch here is that H_3 is data-instigated (that is, created after looking at the data), and the researcher faces choice of a 'post-data prior' involving both $\underline{\pi}_3$ and any parameters unrestricted under H_3. However, the need for sensitivity analysis in public research implies the researcher is simply left with the usual task of presenting a variety of mappings from 'interesting' priors to posteriors. It is left to the reader to decide whether the priors are sufficiently plausible to warrant serious consideration of the data instigated hypothesis. Priors that have been contaminated by data can be presented as such – as always it remains for the reader to assess their plausibility.

Regression

To illustrate the preceding discussion, consider the standard normal linear regression model with fixed regressors X yielding likelihood function $\mathscr{L}(\theta;y) = \varphi_T(y|X\beta, \sigma^2 I_T)$, where $\theta = [\beta', \sigma^{-2}]'$ and β is K × 1 parameter of interest. Working in terms of the precision σ^{-2}, the conjugate normal-gamma prior is

$$f(\beta, \sigma^{-2}) = \Phi_K(\beta|\underline{b}, \sigma^2 Q)\gamma(\sigma^{-2}|\underline{s}^2, \underline{\nu}), \quad (17)$$

where $\boldsymbol{\gamma}(\boldsymbol{\sigma}^{-2}|\underline{\mathbf{s}}^2, \nu) = [2 = \underline{\nu}\underline{\mathbf{s}}^2)^{\nu/2}\Gamma(\underline{\nu} = 2)]^{-1}$ $(\boldsymbol{\sigma}^{-2})^{(\underline{\nu}-2)/2}\exp\left[-\left(\frac{\underline{\nu}\underline{\mathbf{s}}^2}{2}\right)\boldsymbol{\sigma}^{-2}\right]$ is a gamma density with mean $\underline{\mathbf{s}}^{-2}$ and variance $\frac{2}{\underline{\nu}}\underline{\mathbf{s}}^4$, $\Gamma(\cdot)$ denotes the gamma function, b is a K × 1 vector, $\underline{\mathbf{Q}}$ is a K × K positive definite matrix, $\underline{\nu} > 0$, and $\underline{\mathbf{s}} > 0$.

It is the straightforward to show that (5) implies the normal-gamma posterior distribution

$$f(\beta, \boldsymbol{\sigma}^{-2}|\mathbf{y}) = \boldsymbol{\varphi}_K(\beta|\overline{\mathbf{b}}, \boldsymbol{\sigma}^2\overline{\mathbf{Q}})\gamma(\boldsymbol{\sigma}^{-2}|\overline{\mathbf{s}}^2, \overline{\boldsymbol{\nu}}), \quad (18)$$

Where $\overline{b} = \overline{\mathbf{Q}}\left(\underline{\mathbf{Q}} - 1\underline{b} + \mathbf{X}'\mathbf{X}b\right)$, $\overline{\mathbf{Q}} = \left(\underline{\mathbf{Q}}^{-1}\right)\mathbf{X}'$ $\mathbf{X})^{-1}$, $\overline{\nu} = \underline{\nu} + T$ and $\overline{\nu}\overline{s}^{\,2} = \underline{\boldsymbol{\nu}}\,\underline{\mathbf{s}}^2 + (\mathbf{y} - \mathbf{X}b)'$ $(\mathbf{y} - \mathbf{X}b) + (\quad b - \underline{\mathbf{b}})'\left[\underline{\mathbf{Q}} + (\mathbf{X}'\mathbf{X})^{-1}\right]^{-1}(b - \underline{\mathbf{b}})$. The marginal posterior distribution of β is the multivariate-t density

$$f(\beta|\mathbf{y}) = \left[\frac{\pi^{K/2}\Gamma\left(\frac{\overline{\nu}}{2}\right)}{\overline{\nu}^{\overline{\nu}/2}\Gamma\left(\frac{\overline{\nu}+K}{2}\right)}\right]^{-1}$$
$$|\overline{\mathbf{s}}^2\overline{\mathbf{Q}}|^{-1/2}\left[\overline{\nu} + (\beta - \overline{b})'\left(\overline{\mathbf{s}}^2\overline{\mathbf{Q}}\right)^{-1}(\beta - \overline{b})\right]^{-(\overline{\nu}+K)/2} \quad (19)$$

with mean $\overline{\mathbf{b}}$ (if $\overline{\nu} > 1$), variance $\left(\frac{\overline{\nu}}{\overline{\nu}-1}\right)|\overline{\mathbf{s}}^2\overline{\mathbf{Q}}|$(if $\overline{\nu} > 2$), and ν degrees of freedom. The marginal density of the data can be written

$$f(y) = \pi^{-T/2}\left[\frac{\Gamma[(T + \underline{\nu})/2]}{\Gamma(\underline{\nu}/2)}\right]\left[\frac{\left|\underline{\mathbf{Q}}^{-1}\right|}{\left|\overline{\mathbf{Q}}^{-1}\right|}\right]^{1/2}$$
$$\left[\underline{\mathbf{s}}^2\right]^{\underline{\nu}}/2 \cdot \left[\underline{\mathbf{s}}^2 + (\mathbf{y} - \mathbf{X}b)'(\mathbf{y} - \mathbf{X}b) + \right.$$
$$\left. + (b - \underline{\mathbf{b}})'\mathbf{X}'\mathbf{X} + (b - \overline{b}) + (\underline{b} - \overline{b})'\underline{\mathbf{Q}}^{-1}(\underline{b} - \overline{b})\right]$$

Furthermore, the predictive density of an out-of-sample observation $\tilde{y} \in \tilde{\mathbf{Y}}$ corresponding to the regressors $\tilde{\mathbf{x}}$ is the university t density

$$f(\tilde{\mathbf{y}}|\mathbf{y}) = \left[\frac{\pi^{1/2}\Gamma\left(\frac{\overline{\nu}}{2}\right)}{\overline{\mathbf{s}}^{\overline{\nu}/2}\Gamma\left(\frac{\overline{\nu}+1}{2}\right)}\right]^{-1} \left|\tilde{\mathbf{s}}^2\right|^{-1/2} \left[\overline{\nu} + \tilde{\mathbf{s}}^{-2}\left(\mathbf{y} - \tilde{\mathbf{x}}'\overline{\mathbf{b}}\right)'\left(\mathbf{y} - \tilde{\mathbf{x}}'\overline{\mathbf{b}}\right)\right]^{-(\overline{\nu}+1)/2}, \quad (20)$$

where

$$\tilde{\mathbf{s}}^2 = \tilde{\mathbf{s}}^{-2}\left(1 + \tilde{\mathbf{x}}'\overline{\mathbf{Q}}^{-1}\tilde{\mathbf{x}}\right).$$

Note that no full column rank assumption for X is required for the preceding analysis. This reflects a general result that unidentifiability of a parameter, such as β when rank(X) < K, is not much of a problem for a Bayesian with a proper prior, because the posterior is guaranteed to be proper. There is no 'free lunch', however, because there will exist some quantity η about which no learning occurs, that is, $f(\eta|y) = f(\eta)$. For example, if $Xc = 0$ for some nonzero K × 1 vector c, then the prior and posterior distributions for $\eta = \mathbf{c}'\underline{\mathbf{Q}}^{-1}\beta$ given σ^2 is univariate normal with mean $\mathbf{c}'\underline{\mathbf{Q}}^{-1}\underline{\mathbf{b}}$ and variance $\sigma^2\mathbf{c}'\underline{\mathbf{Q}}^{-1}\mathbf{c}$. Whether lack of updating is a problem depends on whether η is a quantity of interest. Note that η depends on both the nature of the collinearity (through c) and the prior (through) $\underline{\mathbf{Q}}$.

Under weighted squared error loss, the Bayesian point estimate of β is the posterior mean $\overline{\mathbf{b}}$. The matrix weighted average of $\underline{\mathbf{b}}$ and is $\overline{\mathbf{b}}$ is precisely the way a classicist combines two samples from the same distribution: a fictitious sample yielding an OLS estimate $\underline{\mathbf{b}}$ with $\mathrm{Var}(\underline{\mathbf{b}}|\sigma^2) = \sigma^2\underline{\mathbf{Q}}$, and an actual sample yielding the OLS estimate b with $\mathrm{Var}(b|\sigma^2) = \sigma^2(X'X)^{-1}$. Elliptical HPD regions for β can be formed using (21). Bayes factors for hypothesis tests involving restrictions on β can be formed from versions of marginal likelihood (22). Finally, under quadratic loss the Bayesian point prediction of $\tilde{\mathbf{y}}$ is $\tilde{\mathbf{y}}_* = \tilde{\mathbf{x}}'\overline{\mathbf{b}}$ and forecast intervals can be obtained directly from the predictive distribution (23).

The standard 'noninformative' prior is $f(\beta,\sigma^{-2}) \propto \sigma^2$, which, unlike the conjugate case, is predicated on the independence of prior beliefs concerning β and σ^{-2}. For this prior, under weighted squared error loss, the Bayesian point estimate of β is the OLS estimate b. HPD regions are numerically identical to frequentist confidence regions of the same level. Under quadratic loss the Bayesian point prediction of $\tilde{\mathbf{y}}$ is $\tilde{\mathbf{y}}_* = \tilde{\mathbf{x}}'\overline{\mathbf{b}}$ and forecast intervals are numerically identical to frequentist forecast intervals of the same level. Bayes factors, however, are not well defined in this case since the prior is improper, and as a result the Bayes factor involves a ratio of arbitrary constants. One class of alternatives in this case are the *intrinsic Bayes factors*, proposed by Berger and Pericchi (1996), which sometimes correspond to actual Bayes factors for particular proper priors known as *intrinsic priors*.

Conclusion

The coherence of the Bayesian approach contrasts sharply with the conventional statistical methods which sometimes advocate negative estimators of positive quantities to ensure unbiasedness, and confidence intervals which may be null or consist of the whole parameter space. Furthermore, Bayesian methods are completely general and do not require usual regularity conditions, asymptotics, sufficient statistics of finite dimension, or pivotal quantities.

There are now a number of textbook sources for Bayesian econometrics. Bayesian econometrics textbooks started with the major contribution of Zellner (1971). While not a textbook as such, Leamer (1978) remains a transparent introduction to Bayesian thinking. Poirier (1995) provides an intermediate level comparison of Bayesian and frequentist reasoning. More recently, Bauwens et al. (1999), Koop (2003), Koop et al. (2007), Lancaster (2004), and Geweke (2005) have covered extensively the statistical models of direct interest to economists. These four texts also serve as excellent introductions to modern computational techniques. Finally, Koop et al. (2006) provides extensive solved Bayesian exercises.

See Also

- ▶ Bayesian Statistics
- ▶ Bayesian Time Series Analysis
- ▶ Markov Chain Monte Carlo Methods

Bibliography

Bauwens, L., M. Lubrano, and J.-F. Richard. 1999. *Bayesian inference in dynamic econometric models*. Oxford: Oxford University Press.

Berger, J.O., and L.R. Pericchi. 1996. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91: 109–122.

Berger, J.O., and R.L. Wolpert. 1988. *The likelihood principle*, 2nd ed. Hayward: Institute of Mathematical Statistics.

Bernardo, J.M. 1979. Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B* 41: 113–147.

Bernardo, J.M., and A.F.M. Smith. 1994. *Bayesian theory*. New York: Wiley.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2003. *Bayesian data analysis*, 2nd ed. New York: Chapman & Hall.

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

Jeffreys, H. 1961. *Theory of probability*, 3rd ed. London: Oxford University Press.

Kadane, J.B., and L.J. Wolfson. 1998. Experiences in elicitation. *Statistician* 47: 3–19.

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. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91: 1343–1370.

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

Koop, G., D.J. Poirier, and J. Tobias. 2007. Bayesian econometric methods. In *Econometrics exercises series*, vol. 7, ed. K. Abadir, J. Magnus, and P.C.B. Phillips. Cambridge: Cambridge University Press.

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

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

Leamer, E.E. 1982. Sets of posterior means and bounded variance priors. *Econometrica* 50: 725–736.

Poirier, D.J. 1988. Frequentist and subjectivist perspectives on the problems of model building in economics (with discussion). *Journal of Economic Perspectives* 2(1): 121–170.

Poirier, D.J. 1995. *Intermediate statistics and econometrics: A comparative approach*. Cambridge, MA: MIT Press.

Press, S.J., and J.M. Tanur. 2001. *The subjectivity of scientists and the Bayesian approach*. New York: Wiley.

Tierney, L., and J.B. Kadane. 1986. Accurate approximations for posterior moments and marginal posterior densities. *Journal of the American Statistical Association* 81: 82–86.

Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.

Bayesian Inference

Arnold Zellner

Bayesian inference is a mode of inductive reasoning that has been used in many sciences, including economics. Bayesian inference procedures are available to evaluate economic hypotheses and models, to estimate values of economic parameters and to predict as yet unobserved values of variables. In addition, Bayesian inference procedures are useful in solving many decision problems including economic control and policy problems, firms' and consumers' stochastic optimization problems, portfolio problems, experimental design problems, etc. Many examples of these uses of Bayesian inference procedures are provided in Jeffreys (1967), De Groot (1970), Zellner (1971), Box and Tiao (1973), Leamer (1978), Boyer and Kihlstrom (1984), and Berger (1985).

A distinctive feature of Bayesian inference procedures is that they permit investigators to use both sample and prior information in a logically consistent manner in making inferences. This is important since prior information is widely used by Bayesian and non-Bayesian workers in making inferences. Bayes' Theorem, sometimes referred to as the Principle of Inverse Probability, serves as a fundamental learning model in the Bayesian approach. Initial or prior information is combined with current sample information by use of Bayes' Theorem to produce a 'post-data' or 'posterior distribution' that incorporates both prior and sample information. In this way prior or initial views are transformed by use of Bayes' Theorem to post-data views, a transformation that is a key, operational learning process.

Thomas Bayes, an 18th-century British Presbyterian minister, is usually given credit for solving the famous 'inverse probability problem', stated by Bayes (1763) as follows: '*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies

somewhere between any two degrees of probability that can be named'. The solution, published two years after Bayes' death, was arrived at by an ingenious geometrical argument. (See Stigler (1983) for further considerations regarding the origins of the solution.) Note that Bayes' inverse problem is fundamentally different from those encountered in games of chance, for example coin flipping, in which the probabilities of outcomes are known and the probabilities of various outcomes must be calculated. These are problems in *direct* probability; for example, calculate the probability of observing five heads in six flips of a fair coin. In Bayes' *inverse* probability problem, five heads in six flips of a coin are observed and what must be calculated or inferred is the chance that the probability of a head on a single flip lies in a given interval, say 0.4–0.7. Thus the probability of a head on a single toss is unknown and must be inferred from the outcomes. The modern solution, due to Laplace (see Molina 1940) will be presented below. It is clear that the inverse problem is typical of scientific problems in which we observe data or outcomes and must infer the probabilistic mechanism or model that probably produced them. Cox (1961) and Jaynes (1984) provide fundamental analysis justifying Bayes' Theorem as a central tool in inductive reasoning.

Since Bayes' essay was published in 1763, Laplace (1820), Edgeworth (1928), Jeffreys (1967, 1973), de Finetti (1970), Wald (1950), Savage (1954), Good (1950, 1965), Lindley (1965, 1971), and many others have contributed to the development of Bayesian analysis and applications of it to many scientific estimation, prediction, testing, and other problems. In what follows, an overview of these developments will be presented and illustrated with analyses of selected problems.

Estimation Problems

Bayes' Theorem plays a central role in estimation problems. Let $\mathbf{y}$ denote a vector of observations contained in a sample space $\boldsymbol{R}\mathbf{y}$ and $\boldsymbol{\theta}$ a vector of parameters contained in a parameter space $\boldsymbol{\Theta}$. Given initial information I_0 let $p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{I}_0)$ be the joint probability density function (pdf) for $\mathbf{y}$ and $\boldsymbol{\theta}$. Then

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{I}_0) &= p(\boldsymbol{\theta}|\boldsymbol{I}_0)p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{I}_0) \\ &= p(\mathbf{y}|\boldsymbol{I}_0)p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{I}_0) \end{aligned} \tag{1}$$

where $p(\cdot|\cdot)$ is a generic symbol for a pdf labelled by its argument. Then from (1),

$$\begin{aligned} p\left(\boldsymbol{\theta}\middle|\mathbf{D}\right) &= p\left(\boldsymbol{\theta}\middle|I_0\right)p\left(\mathbf{y}\middle|\boldsymbol{\theta}, I_0\right)\Big/p\left(\mathbf{y}\middle|I_0\right) \\ &\propto p\left(\boldsymbol{\theta}\middle|I_0\right)p\left(\mathbf{y}\middle|\boldsymbol{\theta}, I_0\right) \end{aligned} \tag{2a}$$

where $D \equiv (\mathbf{y}, \boldsymbol{I}_0)$, the sample and prior information and '$\propto$' denotes 'is proportional to'. The result in (2a) is Bayes' Theorem where $p(\boldsymbol{\theta}|\boldsymbol{D})$ is the posterior pdf for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{I}_0)$ is the prior pdf for $\boldsymbol{\theta}$ and $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{I}_0)$ is the pdf for $\mathbf{y}$ given $\boldsymbol{\theta}$ and $\boldsymbol{I}_0$, which when viewed as a function of $\boldsymbol{\theta}$ is the likelihood function. Thus Bayes' Theorem can be stated as

$$\begin{aligned} \text{Posterior pdf} \propto\ &(\text{Prior pdf}) \\ &\times (\text{Likelihood function}) \end{aligned} \tag{2b}$$

with the factor of proportionality being a normalizing constant.

In (2), $p(\boldsymbol{\theta}|\boldsymbol{I}_0)$, the prior pdf, represents information about possible values for $\boldsymbol{\theta}$ *prior to observing y*. The information in the observations $\mathbf{y}$ is incorporated in the likelihood function and (2) transforms the information in the prior pdf and the likelihood function into a posterior pdf for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{D})$, that is used to make inferences about the possible values of the elements of $\boldsymbol{\theta}$.

It is seen from (2) that the likelihood function plays an important role in Bayes' Theorem in summarizing the sample information. According to the likelihood principle, the likelihood function contains all the sample information and thus no sample information is disregarded when the likelihood function is employed. If there is uncertainty regarding the likelihood function's form, various forms can be considered as explained below and the sample and prior information can be employed in a Bayesian fashion to help resolve the uncertainty.

In the Bayesian approach to inference prior information about the possible values of $\boldsymbol{\theta}$ is formally and explicitly introduced by use of a prior pdf, $p(\boldsymbol{\theta}|\boldsymbol{I}_0)$ in (2). If little prior information is available, a 'diffuse' or 'non-informative' prior pdf is employed, that is one that contains little information about the possible values of $\boldsymbol{\theta}$. On the other hand, if prior information about the possible values of $\boldsymbol{\theta}$ is available, an 'informative' prior pdf would be employed. Prior information may be derived from past studies, economic theory, etc. For example, subject matter considerations and past studies may indicate that a parameter's value falls between zero and one and a prior pdf reflecting this restriction on the range of the parameter would be employed. This is but one type of prior information that may be available and can be incorporated in analyses by use of Bayes' Theorem.

To illustrate the use of Bayes' Theorem in estimation, several simple, important problems will be analysed.

Example 1: Normal Mean with Normal Prior Assume that the observations $y_i = 1, 2, \ldots, n$ have been independently drawn from a normal distribution with unknown mean $\theta, -\infty < \theta < \infty$, and known variance, $\sigma^2 = \sigma_0^2$. The likelihood function is

$$p(\mathbf{y}|\boldsymbol{\theta}, \sigma^2 = \sigma_0^2) = (2\pi\sigma_0^2)^{n/2}\exp\left\{\sum_{i=1}^{n}(y_i - \theta)^2\right\} \propto \exp\left\{-n(\theta - \overline{y})^2/2\sigma_0^2\right\},$$

where

$$\overline{y} = \sum_{i=1}^{n} y_i/n$$

is the sample mean and

$$\sum_{i=1}^{n}(y_i - \theta)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\theta - \overline{y})^2$$

has been employed. Further, assume that prior information regarding θ's possible values is well represented by a normal prior pdf with mean m and variance v, i.e. $p(\theta|m, \upsilon) = (2\pi\upsilon)^{-1/2}\exp\left\{-(\theta - m)^2/2\upsilon\right\}$. Then using Bayes' Theorem in (2), the posterior pdf for θ is

$$\begin{aligned} p(\theta|D) &\propto (\text{prior pdf}) \times (\text{likelihood function}) \\ &\propto \exp\left\{-\left[(\theta - m)^2/\upsilon + n(\theta - \overline{y})^2/\sigma_0^2\right]/2\right\} \\ &\propto \exp\left\{-(\theta - \overline{\theta})^2/2\tau\right\}^2 \end{aligned} \tag{3}$$

a normal pdf with mean $\overline{\theta}$ and variance τ^2 given by

$$\overline{\theta} = (h_0 m + h\overline{y})/(h_0 + h) \tag{4}$$

$$\tau^2 = 1/(h_0 + h) \tag{5}$$

where $h_0 = 1/\upsilon$, the prior precision and $h = n/\sigma_0^2$, the sample precision. It is seen that the posterior mean $\overline{\theta}$ is a weighted average of the prior mean m and the sample mean $\overline{y}$ with the prior precisions, $h_0 = 1/v$ and $h = n/\sigma_0^2$ as weights. As the prior variance $v \to \infty$, that is the prior pdf is very spread out reflecting little information about θ's value, the posterior mean, $\overline{\theta} \to \overline{y}$, the sample mean and the posterior distribution approaches a normal pdf with mean $\overline{y}$ and variance σ_0^2/n Also, as n grows large, the posterior pdf approaches a normal pdf with mean $\overline{y}$ and variance σ_0^2/n For finite υ and n, the normal distribution in (3) can be employed to make probability statements regarding θ's possible value. For example, the posterior probability that $a < \theta < b$ is given by

$$\begin{aligned} \Pr(a < \theta < b|D) &= \int_a^b p(\theta|D)d\theta \\ &= F(b|D) - F(a|D) \end{aligned} \tag{6}$$

where $F(\cdot|D)$ is the cumulative posterior normal distribution associated with (3).

Example 2: Binomial Trials Assume that $\theta, 0 \leq \theta \leq 1$, is the probability of 'success' on a given trial

and that n independent trials yield r successes and $n - r$ failures. The likelihood function is given by

$$\binom{n}{r}\theta^r(1-\theta)^{n-r}$$

Further assume that prior information regarding θ's possible value is well represented by a beta pdf with parameters a and b, that is $p(\theta|a,b) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$ with $0 \le 0 \le 1, a, b > 0$ and where

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta,$$

the beta function. Then the posterior pdf for θ is

$$p(\theta|D) \propto \theta^{r+1a-1}(1-\theta)^{n-r+b-1} \quad 0 \le \theta \le 1 \qquad (7)$$

which is in the beta form with parameters $a' = r + \mathrm{a}$ and $b' = n - r + b$. Thus the normalized posterior pdf is $p(\theta|D) = \theta^{a'-1}(1-\theta)^{b'-1}/B(a',b')$ With the posterior pdf in (7), it is possible to compute the posterior probability that $c_1 < \theta < c_2$, where c_1 and c_2 are any given numbers in the closed interval zero and one, as follows

$$\Pr(c_1 < \theta < c_2|D) = \int_{c_1}^{c_2} p(\theta|D)\mathrm{d}\theta \qquad (8)$$

The integral in (8) can be evaluated using tables of the incomplete beta function or by numerical integration and is a solution to Bayes' inverse probability problem stated earlier. Note that if $a = b = 1$, the prior for θ is uniform over the interval zero to one, the Bayes–Laplace rule for representing little prior information about θ's value – see Jeffreys (1967, pp. 123–125) and Geisser (1984) for further discussion of this rule and other rules for representing knowing little about a binomial parameter.

In the two examples analysed above there were sufficient statistics, $\bar{y}$ for the normal mean problem and r for the binomial problem. It was the case that the posterior distributions were functions of these simple sufficient statistics. This is a general property of Bayesian analyses that is simply shown. Let $\mathbf{t}' = (t_1 + t_2, \ldots, t_m)$ be a vector of sufficient statistics. Then $p(\mathbf{y}|\theta) = h(\mathbf{y})p(\mathbf{t}|\theta)$, where $h(\mathbf{y})$ is a function of just the data $\mathbf{y}$ and Bayes' Theorem in (2) yields

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{D}) &\propto p(\boldsymbol{\theta}|\boldsymbol{I}_0)p(\mathrm{y}|\boldsymbol{\theta}, \boldsymbol{I}_0) \\ &\propto p(\boldsymbol{\theta}|\boldsymbol{I}_0)p(\mathbf{t}|\boldsymbol{\theta}, \boldsymbol{I}_0) \end{aligned} \qquad (9)$$

Thus $p(\theta|D)$ depends on the data just through $\mathbf{t}$, the vector of sufficient statistics.

Further, in both examples analysed above, the prior distributions' forms were in the same form as the likelihood function. When this is the case, the prior distribution is said to have a 'natural conjugate' form – see, e.g. Raiffa and Schlaifer (1961) for further discussion of natural conjugate prior distributions.

Another property of Bayes' Theorem that is quite useful and appealing is that it can be applied sequentially to data sets with results that are identical to what is obtained by an application to an entire data set. To illustrate, consider two independent data vectors, $\mathbf{y}_1$, with pdf $p(\mathbf{y}_1|\theta)$ and $\mathbf{y}_2$ with pdf $p(\mathrm{y}_2|\theta, I_0)$ If $p(\theta|I_0)$ is the prior pdf, then the posterior pdf is

$$p(\boldsymbol{\theta}|D) \propto p(\theta|\boldsymbol{I}_0)p(\mathbf{y}_1|\boldsymbol{\theta}, \boldsymbol{I}_0)p(\mathbf{y}_2|\boldsymbol{I}_0) \qquad (10)$$

Where $D = (\mathbf{y}_1, \mathbf{y}_2, \boldsymbol{I}_0)$. If we analyse the data sets sequentially, $p_1(\theta|D_1) \propto p(\theta|I_0)p(\mathbf{y}_1|\theta, \boldsymbol{I}_0)$ is the posterior pdf based on $D_1 = (\mathbf{y}_1, \boldsymbol{I}_0)$. If $p_1(\theta|D_1)$ is employed as a prior pdf for the analysis of the data set $\mathbf{y}_2$, the posterior pdf is $p(\theta|D) \propto p_1(\theta|D_1)p(\mathbf{y}_2|\theta, \boldsymbol{I}_0)$ which is just the same as (10). Thus the same posterior pdf is obtained by proceeding sequentially as by proceeding as shown in (10).

When a vector of parameters θ is involved in Bayes' Theorem in (2), marginal and conditional posterior pdfs are of interest. Let θ be partitioned as $\theta' = (\theta_1', \theta_2')$ and suppose that interest centres on θ_1. For example θ_2 may be a vector of nuisance parameters that are of little interest to an investigator. The following integration can be performed analytically or numerically to obtain the marginal posterior pdf for θ_1, denoted by $p(\theta_1|D)$,

$$p(\boldsymbol{\theta}_1|D)\int_{\Theta_2} p(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|D)\mathrm{d}\boldsymbol{\theta}_2 \qquad (11a)$$

where Θ_2 is the region containing θ_2. The capability of integrating out nuisance parameters is an extremely important property of the Bayesian approach. Further, writing $p(\theta_1,\theta_2|D) = p(\theta_1|\theta_2,D)p(\theta_2|D)$ where $p(\theta_1|\theta_2,D)$ is the conditional posterior pdf for θ_1 given θ_2 and $p(\theta_2|D)$ is the marginal posterior pdf for θ_2, the integral in (11a) can be expressed as

$$p(\boldsymbol{\theta}_1|D) = \int_{\Theta_2} p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2,D)p(\boldsymbol{\theta}_2|D)d\boldsymbol{\theta}_2 \qquad (11b)$$

Thus the marginal pdf, $p(\theta_1|D)$ can be expressed as an average of conditional posterior pdfs, $p(\theta_1|\theta_2,D)$ with the marginal posterior pdf $p(\theta_2|D)$ as the weight function.

The conditional posterior pdf, $p(\theta_1|\theta_2,D)$ is very important in performing sensitivity analyses. That is, $p(\theta_1|\theta_2,D)$ can be computed for various assigned values for θ_2 to determine how sensitive inferences about θ_1 are to what is assumed about θ_2. See Zellner (1971) and Box and Tiao (1973) for many examples of such sensitivity analyses. For example, θ_2 might be an autocorrelation parameter representing a possible departure from independence and θ_1 a vector of regression coefficients. How a departure from independence affects inferences about regression coefficients can be assessed using conditional posterior distributions.

The large sample properties of posterior distributions is also of interest. Under relatively mild conditions, it has been shown in the literature that as the sample size grows, posterior distributions assume a normal shape with mean approximately equal to the maximum likelihood (ML) estimate and covariance matrix equal to the inverse of the matrix of second derivatives of the log-likelihood function with respect to the parameters evaluated at the ML estimates. Jeffreys (1967, p. 193) views this result as a Bayesian justification for the ML estimate in large samples. For proofs of the asymptotic normality of posterior distributions, see Jeffreys (1967), Heyde and Johnstone (1979), and Hartigan (1983). Heyde and Johnstone (1979) show that when the observations are independently and identically distributed, the conditions needed to prove the asymptotic normality of posterior distributions are identical to those needed to prove asymptotic normality of ML estimators. However, when observations are stochastically dependent, as in time series problems, they show that the conditions needed for asymptotic normality of posterior distributions are simpler and more robust than those needed for proving asymptotic normality of ML estimators.

In summary, Bayes' Theorem provides the complete, finite sample posterior pdf for parameters appearing in all kinds of econometric models. These posterior distributions can be employed to make probability statements about parameters' possible values – see e.g. (6) above for an example. If nuisance parameters are present, they can be integrated out of the joint posterior pdf to obtain a marginal posterior pdf for parameters of interest as shown in (11). Further, if the sample size is large, posterior pdfs assume a normal shape in general with a mean approximately equal to the ML estimate. However, if the sample size is not large, the ML estimate is often not a good approximation to the posterior mean and posterior pdfs' forms are usually non-normal – see Zellner and Rossi (1984) for illustrations of these points using logit models. While the complete posterior pdf is generally available in Bayesian analyses, often interest centres on obtaining an optimal point estimate for a parameter. The Bayesian solution to the problem of point estimation is presented below.

Bayesian Point Estimation

Given a posterior pdf for $\theta \subset \Theta, p(\theta|D)$ derived using Bayes' Theorem in (2), an estimate of θ, say $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(D)$, where $D = (\mathbf{y}, \boldsymbol{I}_0)$, the sample and prior information, is desired. Some measure of central tendency relating to the posterior pdf, say the mean, modal value or median might be used as a point estimate. However, if the posterior pdf is asymmetric, these measures of central tendency will differ and the problem of choice among them

remains. When a loss function, $L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)$, is available, this problem can be solved by choosing the value of $\widehat{\boldsymbol{\theta}}$ that minimizes expected loss and such a value is in Bayesian point estimate. Explicitly, the problem to be solved in Bayesian point estimation is min $EL\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)$ with respect to $\widehat{\boldsymbol{\theta}}$, or

$$\min_{\widehat{\theta}} \int_{\Theta} L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right) p(\boldsymbol{\theta}|D)\mathrm{d}\boldsymbol{\theta} \tag{12}$$

The solution to the minimization problem (12), denoted by $\widehat{\boldsymbol{\theta}}^{*}$ is the Bayesian point estimate. Note that $-L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)$ can be interpreted as a utility function and thus (12) is equivalent to choosing a value for $\widehat{\boldsymbol{\theta}}$ that maximizes expected utility. Given the form of $L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)$ the problem in (12) can be solved analytically or by numerical integration techniques. Below, solutions will be presented for some widely used loss functions.

Quadratic Loss Functions

Let $L\left(\theta,\widehat{\theta}\right) = c_1\left(\widehat{\theta}-\theta\right)^2$ where c_1 is a given positive constant. Then

$$EL = c_1 E\left(\theta-\widehat{\theta}\right)^2 = c_1 E\left[\theta-\overline{\theta}-\left(\widehat{\theta}-\overline{\theta}\right)\right]^2 = c_1\left[E(\theta-\overline{\theta})^2 + \left(\widehat{\theta}-\overline{\theta}\right)^2\right],$$

where $\overline{\theta} = E(\theta|D)$ the posterior mean of θ. Then the value of $\widehat{\boldsymbol{\theta}}$ that minimizes expected loss is $\widehat{\theta}^{*} = \overline{\theta}$, the posterior mean. This is a very general result applicable to all kinds of point estimation problems for which the above 'squared error' loss function is appropriate. For example, in the normal mean problem analysed above, the optimal point estimate relative to the above squared error loss function is the posterior mean given in (4). For the Binomial Trials problem with posterior pdf given in (7) the optimal point estimate relative to squared error loss is the mean of the posterior pdf, namely, $\widehat{\theta}^{*} = (r+a)/(n+a+b)$.

If θ is a vector of parameters and if the loss function is $L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right) = \left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)Q\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$, where Q is a given positive definite symmetric matrix, then

$$\begin{aligned} EL &= E\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)' Q\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \\ &= E\left[\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}-(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})\right]' Q\left[\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}-(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})\right] \\ &= \left(\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\right)' Q\left(\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\right) + E(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})' Q(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}) \end{aligned} \tag{13}$$

where $\overline{\boldsymbol{\theta}} = E\theta|D$ is the posterior mean of θ. From the last line of (13), it is clear that the value of $\widehat{\theta}$ that minimizes expected loss is $\widehat{\theta}^{*} = \overline{\boldsymbol{\theta}}$, the posterior mean. Thus for multiparameter point estimation problems employing a quadratic loss function, the posterior mean is an optimal point estimate in terms of minimizing posterior expected loss.

Absolute Error Loss Functions

If the loss function is $L\left(\theta,\widehat{\theta}\right) = c_2|\widehat{\theta}-\theta|$, = where c_2 is a given positive constant, the value of $\widehat{\theta}$ that minimizes expected loss is the median of the posterior pdf for θ, $p(\theta|D)$. With, $a \le \theta \le b$, posterior expected loss is:

$$\begin{aligned} EL(\theta,\theta) &= c_2\int_a^b |\widehat{\theta}-\theta|\, p(\theta|D)\mathrm{d}\theta \\ &= c_2\left[\int_a^{\widehat{\theta}}\left(\widehat{\theta}-\theta\right)p(\theta|D)\mathrm{d}\theta + \int_{\widehat{\theta}}^{b}\left(\theta-\widehat{\theta}\right)p(\theta|D)\mathrm{d}\theta\right] \end{aligned}$$

Then

$$\frac{\mathrm{d}EL\left(\theta,\widehat{\theta}\right)}{\mathrm{d}\widehat{\theta}} = c_2\left[F\left(\widehat{\theta}|D\right) - 1 + F\left(\widehat{\theta}|D\right)\right] \tag{14}$$

where $F\left(\widehat{\theta}|D\right) = \int_a^{\widehat{\theta}} p(\theta|D)\mathrm{d}\theta$ is the cumulative posterior distribution function. The value of $\widehat{\theta}$ that sets (4) equal to zero is $\widehat{\theta}$ = median of the posterior pdf and this is the value that minimizes expected loss since $\mathrm{d}^2EL/\mathrm{d}\widehat{\theta}^2$, evaluated at the median is strictly positive. Thus for an absolute error loss function, the posterior median is an optimal point

estimate. For the normal mean problem analysed above, the posterior pdf in (3) is normal and hence the median is equal to the mean, given in (4), since the normal posterior pdf is symmetric. For asymmetric posterior pdfs, such as that shown in (7), the median will not be equal to the mean.

Zero–One Loss Functions

If loss is equal to zero as $\widehat{\theta}-\theta$ approaches zero and is equal to one for $|\widehat{\theta}-\theta|\neq 0$, then the modal value of the posterior pdf is the value of $\widehat{\theta}$ that minimizes expected loss – for a proof, see, e.g. Blackwell and Girshick (1954). Thus with a zero–one loss function, the modal value of (7), $\widehat{\theta}^{*}=(r+a-1)/(n+a+b-2)$ is optimal. Note that this value differs from the posterior mean $(r+a)/(n+a+b)$ that is optimal for a squared error loss function.

Asymmetric LINEX Loss Function

Let the loss function be given by

$$L\left(\widehat{\theta}-\theta\right)=b\left[e^{a\left(\widehat{\theta}-\theta\right)}-a\left(\widehat{\theta}-\theta\right)-1\right],\qquad b>0\quad a\neq 0 \tag{15}$$

a class of asymmetric loss functions introduced and used by Varian (1975). For $\widehat{\theta}-\theta=0$, loss is zero and when $a>0$, loss rises almost exponentially for $\widehat{\theta}-\theta>0$ and approximately linearly when $\widehat{\theta}-\theta<0$. The reverse is true when $a<0$. Posterior expected loss is given by

$$EL=b\left[e^{a\widehat{\theta}}Ee^{-a\theta}-a\left(\widehat{\theta}-E\theta\right)-1\right]$$

and the value of $\widehat{\theta}$ that minimizes expected loss is

$$\widehat{\theta}^{*}=-(1/a)\log Ee^{-a\theta} \tag{16}$$

as shown in Zellner (1986). When (16) is evaluated using the normal posterior pdf in (3), the result is

$$\widehat{\theta}^{*}=\overline{\theta}-a\tau^{2}/2 \tag{17}$$

with $\overline{\theta}$ the posterior mean in (4) and τ^2 the posterior variance in (5). It is seen that the optimal point estimate in (17) is less than the posterior mean when $a>0$ and greater than the posterior mean when $a<0$ reflecting the asymmetry of the LINEX loss function in (15).

As the above examples indicate, the Bayesian point estimate is tailored to be optimal relative to the specific loss function that is deemed appropriate. For other loss functions, the problem of minimizing expected loss, shown in (12), can be solved, either analytically or by numerical integration to obtain an optimal Bayesian point estimate. This general procedure is applicable to all estimation problems in econometrics and statistics for which expected loss is finite, that is for which the integral in (12) converges to a finite value.

To appraise the general *sampling properties* of Bayesian estimates, $\widehat{\boldsymbol{\theta}}=\widehat{\boldsymbol{\theta}}(\mathbf{y},\boldsymbol{I}_0)$ is regarded as a random estimator. Relative to a specific loss function, $L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)$, the risk function is given by

$$r(\boldsymbol{\theta})=\int_{R_y}\boldsymbol{L}\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{I}_0)\mathrm{d}\mathbf{y} \tag{18}$$

and the Bayesian estimator is, by definition, the one that minimizes average or Bayes risk (BR),

$$\mathrm{BR}=\int_{\Theta}r(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{I}_0)\mathrm{d}\boldsymbol{\theta} \tag{19}$$

where $p(\boldsymbol{\theta}|\boldsymbol{I}_0)$ is a given prior pdf for $\boldsymbol{\theta}$ that is assumed to be positive over the region $\boldsymbol{\Theta}$. Upon substituting (18) in (19),

$$\mathrm{BR}=\int_{\Theta}\int_{R_y}L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{I}_0)p(\boldsymbol{\theta}|\boldsymbol{I}_0)\mathrm{d}\mathbf{y}\,\mathrm{d}\boldsymbol{\theta} \tag{20a}$$

Using $p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{I}_0)p(\boldsymbol{\theta}|\boldsymbol{I}_0)=p(\boldsymbol{\theta}|D)p(\mathbf{y}|\boldsymbol{I}_0)$, where $D=(\mathbf{y},\boldsymbol{I}_0)$, and interchanging the order of integration, (20a) can be expressed as

$$\mathrm{BR}=\int_{R_y}\left[\int_{\Theta}L\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right)p(\boldsymbol{\theta}|D)\mathrm{d}\boldsymbol{\theta}\right]p(\mathbf{y}|\boldsymbol{I}_0)\mathrm{d}\mathbf{y} \tag{20b}$$

If the integral defining BR converges to a finite value and if $p(\mathbf{y}|\boldsymbol{I}_0) > 0$ over the region Ry, then the value of $\widehat{\boldsymbol{\theta}}$ that minimizes the integral in square brackets in (20b) minimizes BR, i.e. it is the estimator that minimizes Bayes or average risk. Note that the integral in square brackets defines posterior expected loss. Thus the solution to the problem in (12), viewed as an estimator is the estimator that minimizes BR and is by definition the Bayesian estimator. This estimator is admissible since if there were another estimator that had lower risk, $r(\theta)$ given in (18), over Θ, it would have lower BR, a contradiction since the Bayesian estimator, by construction minimizes BR. These and other properties of Bayesian estimators are discussed in De Groot (1970), Berger (1985) and other works on decision theory. Thus Bayesian estimators relative to the loss function and prior pdf used to derive them have very good sampling properties under the condition that BR is finite, a sufficient condition for admissibility of the Bayesian estimator.

To illustrate some of the above concepts, consider estimation of a mean θ in the normal mean problem in Example 1 relative to a squared error loss function $L\left(\theta,\widehat{\theta}\right) = \left(\theta - \widehat{\theta}\right)^2$. As mentioned above, the posterior mean in (4) is the Bayesian estimator for this problem that we write as

$$\overline{\theta} = wm + (1-w)\overline{y} \tag{21}$$

with $w = h_0/(h + h_0)$. Since $\overline{y}$ is normally distributed with mean θ and variance σ_0^2/n, the risk of $\overline{\theta}$ relative to a squared error loss function is

$$\begin{aligned} r(\theta) = E_{\overline{y}}\left(\theta - \overline{\theta}\right)^2 &= \theta^2 - 2\theta E\overline{\theta} + E\overline{\theta}^2 \\ &= \theta^2 - 2\theta[wm + (1-w)E\overline{y}] \\ &\quad + E[wm + (1-w)\overline{y}]^2 r(\theta) \\ &= w^2(\theta - m)^2 + (1-w)^2\sigma_0^2/n \end{aligned} \tag{22}$$

It is clear that $r(\theta)$ is smallest when $\theta = m$, the prior mean. To compute BR, we average $r(\theta)$ using the normal prior for θ with mean m and variance $v = 1/h_0$ to obtain

$$\mathrm{BR} = w^2 v + (1-w)^2\sigma_0^2/\boldsymbol{n} = 1/(h_0 + h) \tag{23}$$

where $\sigma_0^2/n = 1/h$ For comparison, the risk function for the sample mean, $\overline{y}$, is

$$r(\theta) = E(\theta - \overline{y})^2 = \sigma_0^2/n \tag{24}$$

On comparing (24) and (22), it is seen that when θ is close to the prior mean, (22) is smaller than (24). The BR of the sample mean is

$$\mathrm{BR} = \sigma_0^2/n = 1/h \tag{25}$$

which is larger than the BR of the posterior mean in (23). This is not surprising since the posterior mean is the estimator that minimizes BR.

Above, proper prior pdfs were employed to obtain Bayesian estimates and estimators. When *improper* prior pdfs are employed to represent vague or little prior information about parameters' values as in Jeffreys (1967) and others' works, posterior pdfs are usually proper and posterior probability statements can be made. Rényi (1970) – see also Hartigan (1983) – has provided an axiom system for probability theory that accommodates improper prior pdfs or unbounded measures and within the context of which Bayes' Theorem remains valid. When such prior pdfs are employed, the solution to the point estimation problem in (12) is termed a generalized Bayes estimate (GBE). Often BR is not finite for GB estimators and they need not in general be admissible. To illustrate, a normal mean problem and a regression problem will be analysed employing diffuse, improper prior pdfs.

Example 3: Normal Mean with an Improper Prior Assume that n observations have been independently drawn from a normal distribution with mean θ and variance σ^2, both of which have unknown values. The likelihood function is given by

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta},\sigma) &\propto \sigma^{-n}\exp\left\{-(\mathbf{y} - \boldsymbol{\iota}\theta)'(\mathbf{y} - \boldsymbol{\iota}\theta)/2\sigma^2\right\} \\ &\propto \sigma^{-n}\exp\left\{-\left[vs^2 + n(\theta - \overline{y})^2\right]/2\sigma^2\right\} \end{aligned} \tag{26}$$

where $\mathbf{y}' = (y_1, y_2, \ldots, y_n), \boldsymbol{\iota}' = (1, 1, \ldots, 1)$ a $1 \times n$ vector with all elements equal to one,

$$\bar{y} = \sum_{i=1}^{n} y_i/n,$$

the sample mean and

$$vs^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ and } v = n - 1.$$

Jeffreys's (1967) diffuse improper prior pdf for this problem is,

$$p(\theta, \sigma) \propto 1/\sigma. \quad \begin{matrix} -\infty < \theta < \infty \\ 0 < \sigma < \infty \end{matrix} \tag{27}$$

That is θ and log σ are assumed independently and uniformly distributed. Since the integral of $p(\theta, \sigma)$ over the range $-\infty < \theta < \infty$ and $0 < \sigma < \infty s$ does not converge to one, the prior pdf is termed 'improper'. See Zellner (1971, 1977) and Berger (1985) for further discussion of (27). On combining the likelihood function in (26) with the prior in (27) by Bayes' Theorem, the result is the joint posterior pdf for θ and σ, namely

$$p(\theta, \sigma|D) \propto \sigma^{-(n+1)} \exp\left\{-vs^2 + n(\theta - \bar{y})^2\right]/2\sigma^2\}. \tag{28}$$

From (28), it is seen that the conditional posterior pdf for θ given σ is normal with mean y and variance σ^2/n. Also by integrating (28) with respect to σ, 0 to ∞, the marginal posterior pdf for θ is

$$p(\theta|D) \propto \left\{vs^2 + n(\theta - \bar{y}^2\right\}^{-(v+1)/2} \tag{29}$$

which is a proper pdf, given $v > 0$, in the univariate Student-t form. That is $t = \sqrt{n}(\theta - \bar{y})/s$ has a standardised Student-t pdf with v degrees of freedom. Also, on integrating (28) with respect to θ, from $-\infty$, to ∞ the result is the marginal posterior for σ,

$$p(\sigma|D) \propto \sigma^{-(v+1)} \exp\left\{-vs^2/2\sigma^2\right\}, \tag{30}$$

a proper pdf in the inverted-gamma form – see, e.g. Zellner (1971) for its properties. Thus even though the improper prior pdf in (27) was employed, the posterior pdfs in (29) and (30) are proper and can be employed to make posterior probability statements about the values of θ and σ.

To obtain an optimal point estimate for θ, assume that a squared error loss function is appropriate, that is $L\left(\theta, \widehat{\theta}\right) = \left(\widehat{\theta} - \theta\right)^2$ Relative to this loss function the value of $\widehat{\theta}$ that minimizes posterior expected loss is the posterior mean of (29) which is $\bar{y}$ for $v > 1$. The risk of $\bar{y}$ relative to the squared error loss function is $E(\bar{y} - \theta)^2 = \sigma^2/n$ since $\bar{y}$ has a normal pdf with mean θ and variance σ^2/n If we try to compute the BR of $\bar{y}$ relative to the improper prior in (27), it is clear that BR is unbounded, that is the integral defining

$$\text{BR} = \int_0^\infty \int_{-\infty}^\infty r(\theta) p(\theta, \sigma|I_0) \mathrm{d}\theta \, \mathrm{d}\sigma$$

diverges. Thus $\bar{y}$, the posterior mean does not minimize BR. A different argument must be used to establish the admissibility of $\bar{y}$ – see e.g. Blyth (1951) and Berger (1985) for proofs of the admissibility of $\bar{y}$.

As regards the posterior pdf for σ in (30), it can be transformed to a posterior pdf for σ^2 by a simple change of variable from σ to $\phi = \sigma^2$ to yield

$$p(\phi|D) \propto \phi^{-(n+2)/2} \exp\left\{-vs^2/2\phi\right\} \quad 0 < \phi < \infty \tag{31}$$

The posterior mean of ϕ is $\overline{\phi} = vs^2/(v-2)$, for $v > 2$ which is optimal relative to a squared error loss function. Also, with respect to a relative squared error loss function, $L\left(\widehat{\phi}, \phi\right) = \left(\widehat{\phi} - \phi\right)^2/\phi^2$, the optimal value of $\widehat{\phi}$ is $\widehat{\phi} = vs^2/(v+2)$ since $EL = \widehat{\phi}^2 E1/\phi^2 - 2\widehat{\phi} E1/\phi + 1$ and the minimizing value of $\widehat{\phi}$ is zero one loss function is $\widehat{\phi}_{m0} = vs^2/(v+2)$ Thus point estimates that are optimal relative to various loss functions are readily obtained. Finally from (30) vs^2/ϕ has a χ_v^2 posterior pdf, a fact that is very useful in making posterior probability statements regarding ϕ's values.

Example 4: Normal Regression Model with a Diffuse Prior Assume that the $n \times 1$ observation vector $\mathbf{y}$ is generated by $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$ where X is an $n \times k$ non-stochastic matrix with rank k, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients with unknown values and $\mathbf{u}$ is an $n \times 1$ vector of disturbance terms assumed independently drawn from a normal pdf with zero mean and finite variance σ^2 with unknown value. The likelihood function under these assumptions is

$$p(\mathbf{y}|X,\boldsymbol{\beta},\sigma) \propto \sigma^{-n}\exp\{-(\mathbf{y}-X\boldsymbol{\beta})'(\mathbf{y}-X\boldsymbol{\beta})/2\sigma^2\}$$
$$\propto \sigma^{-n}\exp\left\{-\left[vs^2 + \left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'X'X\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right]/2\sigma^2\right\} \tag{32}$$

where

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}, vs^2 = \left(\mathbf{y}-X\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-X\widehat{\boldsymbol{\beta}}\right)$$

and $v = n = k$. The diffuse prior pdf that will be employed is

$$p(\boldsymbol{\beta},\sigma|I_0) \propto 1/\sigma \quad -\infty < \beta_i < \infty \quad i = 1,2,\ldots,k \quad 0 < \sigma < \infty \tag{33}$$

That is the elements of $\boldsymbol{\beta}$ and $\log\sigma$ are assumed to be uniformly and independently distributed. Since (32) does not integrate to a constant, it is an improper pdf. However on combining it with the likelihood function in (32) by means of Bayes' Theorem, the resulting joint posterior pdf is,

$$p(\boldsymbol{\beta},\sigma|D) \propto \sigma^{-(n+1)}$$
$$\exp\left\{-\left[vs^2 + \left(\beta-\widehat{\beta}\right)'X'X\left(\beta-\widehat{\beta}\right)\right]/2\sigma^2\right\} \tag{34}$$

From (34), it is seen that the conditional posterior pdf for $\boldsymbol{\beta}$ given σ is normal with mean $\widehat{\boldsymbol{\beta}}$ and covariance matrix $(X'X)^{-1}\sigma^2$. Since σ's value is unknown, this result is not very useful in practice. To get rid of the nuisance parameter σ, (34) is integrated with respect to σ to yield the marginal posterior pdf for β,

$$p(\boldsymbol{\beta}|D) \propto \left\{vs^2 + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'X'X\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right]\}^{-(v+k)/2}\} \tag{35}$$

a posterior pdf that is the multivariate Student-t form–see, e.g. Raiffa and Schlaifer (1961) and Zellner (1971) for its properties. For $v > 1$, the mean of (34) is $E(\boldsymbol{\beta}/D) = \widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathrm{y}$ the least squares quantity and ML estimate. This then is another example wherein a non-Bayesian result has been produced by the Bayesian approach. Further, from (35), the marginal pdf for an element of β, say β_i is in the univariate Student-t form; that is, $\left(\beta_i - \widehat{\beta}_i\right)s_{\widehat{\beta}_i}$ has a univariate Student-t pdf with v degrees of freedom where $\widehat{\beta}_i$ is the ith element of $\widehat{\boldsymbol{\beta}}$ and $s^2_{\widehat{\beta}_i} = m^{ii}s^2$, where m^{ii} is the i-ith element of $(X'X)^{-1}$ Thus posterior probability statements about β_i's value, e.g. $\Pr(\beta_i > 0|D)$ can be made using properties of the univariate Student-t pdf.

Further, on integrating (34) with respect to the elements of $\boldsymbol{\beta}$, the following marginal posterior pdf for σ is obtained

$$p(\sigma|D) \propto \sigma^{-(v+1)}\exp\{-vs^2/2\sigma^2\} \tag{36}$$

a posterior pdf in the 'inverted gamma' form–see, e.g. Raiffa and Schlaifer (1961) and Zellner (1971) for its properties. By a change of variable in (36), the posterior pdf for $\phi = \sigma^2$, the variance is

$$p(\phi|D) \propto \phi^{-(v+2)/2}\exp\{-vs^2/2\phi\} \tag{37}$$

for which it is the case that vs^2/ϕ has a χ^2_v, pdf with $v = n - k$ degrees of freedom. The modal values and moments of (36) and (37) are readily available. Also, posterior probability statements regarding ϕ's possible values can be evaluated using tables of the χ^2_v pdf; that is the [posterior probability that ϕ lies between vs^2/a_2 and vs^2/a_1, given by $\Pr\{vs^2/a_2 < \phi < vs^2/a_1|D\}$, where a_1, $a_2 > 0$ are given constants, can be evaluated by use of χ^2_v tables by noting that the required probability is equal to $\Pr\{a_1 < vs^2/\phi < a_2|D\}$, where vs^2/ϕ has a χ^2_v posterior pdf.

As can be seen from what has been presented above, Bayesian point estimates can be readily computed from posterior pdfs that are optimal relative to loss functions that are deemed appropriate. They reflect both sample and prior

information, as little or as much of the latter as is available. As regards sampling properties of point estimates, some properties of Bayesian estimators have been noted above. The relevance of sampling properties for making inferences from a *given* sample of data has been questioned by some–see e.g. Tiao and Box (1975). Indeed, it is difficult to state which sequence of future samples is most relevant for a given problem. Usually the sequence considered is *identical repetitions* of the process giving the sample data. This sequence is often not the most relevant sequence. However, before the sample data are drawn, it appears relevant to consider possible outcomes, particularly with respect to design of experiments, and it is here that sampling properties of procedures, including point estimation procedures, are most relevant. Once the data are drawn, the researcher's task is to make inferences based on the given sample and prior information.

With this said about point estimation, attention will now be given to interval estimation.

Interval Estimation

Given that a posterior pdf for a parameter θ, $p(\theta|D)$ is available that is unimodal, in interval estimation an interval is sought within which the parameter's value lies with a specified posterior probability, say 0.95. Since such intervals are not unique, it is necessary to impose a condition so that a unique interval is obtained. The condition is that of all intervals with posterior probability $1 - \alpha$, the one selected is the shortest. Formally, the length of the interval, say $b - a$ is minimized Subject to the condition that $\Pr(a < \theta < b|D) = \int_a^b p(\theta|D)d\theta = 1 - \alpha$, the given posterior probability. The solution for this constrained minimization problem is to take the values of a and b such that $p(a|D) = p(b|D)$–see, e.g. Zellner (1971) for a proof. Given that $p(\theta|D)$ is unimodal, the posterior interval with probability content $1 - \alpha$ so computed has posterior densities associated with it that are greater than any other interval with posterior probability $1 - \alpha$ and thus has been called a posterior highest density (PHD) interval.

Example 5: PHD Interval for a Regression Coefficient In Example 4, it was indicated that the posterior pdf for a regression coefficient β_i is in the univariate Student-t form, that is $t_v = \left(\beta_i - \widehat{\beta}_i\right)/s_{\widehat{\beta i}}$ has a univariate Student-t posterior pdf with $v = n - k$ degrees of freedom. Then, with given probability $1 - \alpha$, say 0.95, $\Pr(-c < t_v < c|D) = 1 - \alpha$, where $c > 0$ is obtained from t-tables with v degrees of freedom. Note from the symmetry of the Student-t pdf, $p(-c|D) = p(c|D)$ as required for a PHD interval. Also the event $-c < t_v < c$ is equivalent to $\widehat{\beta}_i - c s_{\widehat{\beta}} < \beta_i < \widehat{\beta}_i + c s_{\widehat{\beta}}$ and thus the posterior probability that β_i is in the given interval $\widehat{\beta}_i + c s_{\widehat{\beta}}$ is 0.95.

Further, a posterior region for the regression coefficient vector can be computed using the following result from (35), a property of the multivariate Student-t pdf, namely,

$$F_{k,v} = \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)' X'X \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)/ks^2 \qquad (38)$$

which has a posterior F distribution with k and v degrees of freedom. When β has two elements, (38) can be employed to compute a confidence region in the form of an ellipse with a given posterior probability, $1 - \alpha$, such that β_1 and β_2 fall within it by choosing a value of $F_{k,v}$, say F_α such that $\Pr(F_{k,v} \le F_\alpha) = 1 - \alpha$.

Bayesian Prediction Procedures

Let $\mathbf{y}_f$ represent a vector of as yet unobserved variables and assume that the pdf for $\mathbf{y}_f$ is $f\left(\mathbf{y}_f|\theta\right)$ where $\boldsymbol{\theta}$ is a vector of parameters with unknown values. The fact that $\boldsymbol{\theta}$ has an unknown value makes it difficult to use $f\left(\mathbf{y}_f|\boldsymbol{\theta}\right)$ to make probability statements about possible values of $\mathbf{y}_f$ However, if a posterior pdf for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|D)$ is available, provided by Bayes' Theorm in (2), then the joint pdf for $\mathbf{y}_f$ and θ is given by $f\left(\mathbf{y}_f|\boldsymbol{\theta}\right)p(\boldsymbol{\theta}|D)$, where D represents past data and prior information. Then the marginal or predictive pdf for $\mathbf{y}_f$, $p\left(\mathbf{y}_f|D\right)$, is given by,

$$p\left(\mathbf{y}_f|D\right) = \int_{\Theta} f\left(\mathbf{y}_f|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|D)\mathrm{d}\boldsymbol{\theta}. \tag{39}$$

From (39) it is seen that the predictive pdf can be interpreted as an average of $f\left(\mathbf{y}_f|\boldsymbol{\theta}\right)$ with $p(\boldsymbol{\theta}|D)$ serving as the weight function. The result in (39) gives the complete predictive pdf for the vector of future values, $\mathbf{y}f$ from which marginal pdfs for particular elements of $\mathbf{y}_f$ can be obtained by integration. Also, moments of the elements of $\mathbf{y}_f$ can be evaluated. As explained below, the mean of the predictive pdf (39) is an optimal point prediction relative to a quadratic loss function. Also, predictive intervals and regions for elements of $\mathbf{y}_f$ can be computed from (39)

As regards point prediction, given a predictive loss function, $L(\mathbf{y}_f, \hat{\mathbf{y}}_f)$ where $\widehat{\mathbf{y}}_f = \widehat{\mathbf{y}}_f(D)$ is some point prediction, an optimal value for $\hat{\mathbf{y}}_f$ is obtained by minimizing expected loss, that is by solving the following problem:

$$\min_{\widehat{\mathbf{y}}_f} \int L\left(\mathbf{y}_f, \widehat{\mathbf{y}}_f\right) p\left(\mathbf{y}_f|D\right) d\mathbf{y}_f. \tag{40}$$

The solution, $\widehat{\mathbf{y}}_f^*$ is the optimal point prediction. For example, for a quadratic loss function, $L\left(\mathbf{y}_f, \widehat{\mathbf{y}}_f\right) = \left(\mathbf{y}_f - \widehat{\mathbf{y}}_f\right)' Q\left(\mathbf{y}_f - \widehat{\mathbf{y}}_f\right)$ where Q is a given positive definite symmetric matrix, expected loss is given by

$$\begin{aligned} &E\left(\mathbf{y}_f - \widehat{\mathbf{y}}_f\right)' Q\left(\mathbf{y}_f - \widehat{\mathbf{y}}_f\right) \\ &= E\left[\mathbf{y}_f - \overline{\mathbf{y}}_f - \left(\widehat{\mathbf{y}}_f - \overline{\mathbf{y}}_f\right)\right]' Q\left[\mathbf{y}_f - \overline{\mathbf{y}}_f - \left(\widehat{\mathbf{y}}_f - \overline{\mathbf{y}}_f\right)\right] \\ &= E\left(y_f - \overline{y}_f\right)' Q\left(y_f - \overline{y}_f\right) + \left(\widehat{y}_f - \overline{y}_f\right)' Q\left(\widehat{y} - \overline{y}_f\right) \end{aligned} \tag{41}$$

where $\overline{\mathbf{y}}_f$ is the mean of the predictive pdf. From (41), it is clear that taking $\widehat{y}_f = \overline{y}_f$ minimizes expected loss. Thus, in general, the mean of a predictive pdf is an optimal point prediction relative to quadratic loss. As in the case of point estimation, if other loss functions are employed, point predictions that are optimal relative to them can be calculated by solving the problem in (40) analytically or numerically. This analysis for absolute error, zero–one and LINEX loss functions is similar to that presented above in connection with point estimation. Also, Bayesian point predictors based on proper prior distributions are admissible and minimize Bayes risk.

To illustrate the calculation of a predictive pdf for the multiple regression model with the posterior pdf for its parameters given in (34), let a future scalar observation, y_f be given by

$$y_f = \mathbf{x}_f'\boldsymbol{\beta} + u_f \tag{42}$$

where $\mathbf{x}_f'$ is a $1 \times k$ given vector and uf is a normal error term with zero mean and variance σ^2. Then noting from (34) that the posterior pdf of β given σ is $N\left[\widehat{\beta}, (X'X)^{-1}\sigma^2\right)$ the conditional distribution of y_f given σ is normal with mean $\mathbf{x}_f'\boldsymbol{\beta}$ and covariance matrix, $\left[1 + \mathbf{x}_f'(X'X)^{-1}\mathbf{x}_f\right]\sigma^2$. On multiplying this conditional predictive pdf for y_f by the posterior pdf for σ, given in (36) and integrating over σ, the result is

$$p\left(y_f|D\right) \propto \left\{vs^2 + \left(y_f - \widehat{y}_f\right)^2/a^2\right\}^{-(v+1)/2} \tag{43}$$

where $\widehat{y}_f = \mathbf{x}_f'\boldsymbol{\beta}, a^2 = 1 + \mathbf{x}_f'(X'X)^{-1}\mathbf{x}_f$ and $v = n - k$, a pdf in the univariate Student-t form with v degrees of freedom with mean $\hat{y}_f$. Thus

$$t_v = \left(y_f - \widehat{y}_f\right)/as \tag{44}$$

has a univariate Student-t pdf with v degrees of freedom. Using (44), a predictive interval for y_f can be computed that has a given probability, say $1 - \alpha$, of including y_f. Such an interval takes the form $\widehat{y}_f \pm c_{\alpha/2}$ as, where $c_{\alpha/2}$ is a constant obtained from tables of the t-distribution. The probability statement associated with this interval is:

$$\begin{aligned} &\Pr\left\{\widehat{y}_f - c_{\alpha/2}as < y_f < \widehat{y}_f + c_{\alpha/2}as|D\right\} \\ &\quad = 1 - \alpha. \end{aligned} \tag{45}$$

Note that y_f is random and the endpoints of the interval are non-random since they depend just on

the given data D. A similar analysis can be performed to obtain the predictive pdf for a vector of future values y_f assumed generated by $\mathbf{y}_f = X_f \boldsymbol{\beta} + \mathbf{u}_f$ when the value of X_f is given.

Above in (42), $\mathbf{x}_f$ was assumed given. If $\mathbf{x}_f$'s value is unknown then the predictive pdf for y_f is given by

$$p\left(\mathrm{y}_f|D\right) = \int p\left(y_f|x_f, D\right) p\left(\mathbf{x}_f|D_1\right) \mathrm{d}\mathbf{x}_f \qquad (46)$$

Where $p\left(x_f|D_1\right)$ is the predictive pdf for $\mathbf{x}_f$, given data D_1. The integration in (46) may be performed analytically or numerically. Further, predictive pdfs can be computed for linear combinations of future values, say $\mathbf{z} = A\mathbf{y}_f$ where A is a given matrix, for time series models–see Broemeling (1985), Monahan (1983) and Zellner (1971); for simultaneous equation models–see Richard (1973), and for many other models–see Aitchison and Dunsmore (1975).

Bayesian Analysis of Hypotheses

Bayesian methods are available for analysing hypotheses about parameters' values and for comparing and choosing between alternative hypotheses or models, be they nested or non-nested. Prior probabilities are assigned to hypotheses or models that reflect the degrees of confidence associated with them and Bayes' Theorem is employed to compute posterior probabilities for them that reflect the information in sample data. This approach differs radically from non-Bayesian testing procedures in which one hypothesis, e.g. the null hypothesis, or one of two models is assumed to be 'true' and a test statistics' distribution, derived under an assumed true null hypothesis or model is employed to 'accept' or 'reject' the assumed true hypothesis or model. For further consideration of these issues see Jeffreys (1967), Jaynes (1984), Kruskal (1978), Leamer (1978), and Zellner (1971, 1984).

To illustrate the Bayesian approach for analysing hypotheses, consider first a scalar parameter θ, say a population mean or a regression coefficient with possible values $-\infty < \theta < \infty$.

Assume that the following two hypotheses are of interest, $H_i : \theta > 0$ and $H_2 : \theta \leq 0$. Given a prior pdf for the parameter, $p(\theta|I_0)$, the prior probability that $\theta > 0$ is given by

$$\Pr(\theta > 0|I_0) = \int_0^\infty p(\theta|I_0)\mathrm{d}\theta \qquad (47)$$

and the prior probability that $\theta \leq 0$ is

$$\Pr(\theta \leq 0|I_0) = \int_{-\infty}^0 p(\theta|I_0)d\theta = 1 - \Pr(\theta > 0|I_0)$$

Then the prior odds for H_1 versus H_2, denoted by K_{12}^0 is

$$K_{12}^0 = \Pr(\theta > 0|I_0)/\Pr(\theta \leq 0|I_0) \qquad (48)$$

These probabilities and K_{12} summarize initial views of the hypotheses H_1 and H_2. If data $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ are observed relating to θ's possible value, Bayes' Theorem in (2) can be employed to compute the posterior pdf for θ, $\mathrm{p}(\theta|D)$, where $D = (y, I_0)$ represents the sample and prior information. Then

$$\Pr(\theta > 0|D) = \int_0^\infty p(\theta|D)\mathrm{d}\theta \qquad (49)$$

and

$$\Pr(\theta \leq 0|D) \int_{-\infty}^0 p(\theta|D)\mathrm{d}\theta \qquad (50)$$

are the posterior probabilities associated with H_1 and H_2, respectively and their ratio, K_{12}, is the posterior odds. The posterior probabilities in (49) and (50) differ from (47) and (48) because the former incorporate the information in the data. This approach can be extended to cases in which θ is a vector, say $\boldsymbol{\theta}' = (\theta_1, \theta_2)$ and hypotheses such as $H_1 : \theta_1 > 0$ and $\theta_2 > 0$; $H_2 : \theta_1 \leq 0$ and $\theta_2 > 0$; $H_3 : \theta_1 > 0$ and $\theta_2 \leq 0$ and H_4: $\theta_1 \leq 0$ and $\theta_2 \leq 0$. In analyses of these four hypotheses, bivariate prior and posterior pdfs for θ_1 and θ_2 can be employed to compute probabilities associated with each of the four hypotheses. These and other

Bayesian Inference, Table 1

Acts	States of world		Expected loss
	H_1 is appropriate	H_2 is appropriate	
Choose H_1	0	L_{12}	$L_{12}\Pr(H_2\|D)$
Choose H_2	L_{21}	0	$L_{21}\Pr(H_1\|D)$
Probabilities	$\Pr(H_1\|D)$	$\Pr(H_2\|D)$	

hypotheses, involving inequality constraints on parameters' values, are easily analysed. The integrals giving probabilities can be evaluated analytically or numerically – see Zellner (1971, pp. 194–200) for an example of this type of analysis relating to a second order autoregressive process.

Above, various probabilities associated with hypotheses have been computed. Given a loss structure, it is possible to choose a hypothesis so as to minimize expected loss. That is, in considering two hypotheses, a two-action-two-state loss structure is shown in Table 1, where L_{12} is the loss incurred when H_1 is selected and H_2 is appropriate, an error of type II, whereas loss L_{21} is incurred if H_2 is chosen when H_1 is appropriate, an error of type I. Given probabilities, $\Pr(H_1|D)$ and $\Pr(H_2|D)$, say computed from (49) and (50), respectively, they can be used to compute expected losses shown in the last column of the table. Then H_1 is chosen if $L_{12}\Pr(H_2|D) < L_{21}\Pr(H_1|D)$ and H_2 otherwise. The condition for choosing H_1 is:

$$1 < L_{21}\Pr(H_1|D)L_{12}\Pr(H_2|D) \tag{51}$$

In the special case of a symmetric loss structure, $L_{12} = L_{21}$, the decision rule in (51) reduces to choose H_1 if $\Pr(H_1|D) > \Pr(H_2|D)$. Note that the decision rule in (51) reflects prior and sample information as well as the loss structure.

In the examples considered above, the hypotheses considered did not involve assigning a specific value to a parameter, e.g. $\theta = 1$ or $\theta = 1$. Since hypotheses such as these are frequently encountered in applied work, it is important to be able to appraise them. Jeffreys (1967, Chs V and VI), who is a pioneer in this area, has provided many Bayesian solutions for such testing problems. In this approach, a hypothesis $H_1 : \theta = 0$ is considered relative to another hypothesis, $H_2 : \theta = \theta_2$ is a given value different from 0. Prior probabilities, Π_1 and Π_2 are assigned to H_1 and H_2, respectively. Assume further that a vector of observations y is available and that $p(y|\theta = \theta)$ is the likelihood function under hypothesis H_1 and $p(y|\theta = \theta_2)$ is the likelihood function under hypothesis H_2. Then Bayes' Theorem is employed to obtain the following posterior odds, K_{12} relating to H_1 versus H_2:

$$\begin{aligned} K_{12} &= (\Pi_1/\Pi_2) \times \{p(\mathrm{y}|\theta = 0)/p(\mathrm{y}|\theta = 0_2)\} \\ &= (\text{Prior Odds}) \times (\text{Bayes' Factor}). \end{aligned} \tag{52}$$

In (52) $\Pi_1/\Pi_2 \frac{-b\pm\sqrt{b^2-4ac}}{2a}$ is the prior odds for H_1 versus H_2 while the Bayes' Factor (BF) is the ratio of likelihood functions, $p(\mathbf{y}|\theta = 0/p(\mathbf{y}|\theta = \theta_2)$. The result in (52) can be regarded as a transformation of the prior odds into a posterior odds reflecting both prior and sample information. The following example illustrates use of (52).

Example 6: Posterior Odds for Two Simple Hypotheses Let $y_i = \theta + \varepsilon_i, i = 1, 2, \ldots, n$ where the ε_i's have been independently drawn from a normal distribution with zero mean and unit variance. Consider two hypotheses about the value of the mean, θ, $H_1 : \theta = 0$ and $H_2 : \theta = \theta_2 > 0$, with prior odds, with prior odds, $\Pi_1/\Pi_2 = 1$. Then the posterior odds are given by $K_{12} = (\Pi_1/\Pi_2)\ p(\mathrm{y}|\theta = 0)/\mathrm{p}(\mathrm{y}|\theta = \theta_2)$, or with $\Pi_1/\Pi_2 = 1$,

$$\begin{aligned} K_{12} &= \exp\left\{-\sum_1^n y_i^2/2\right\}/\exp\left\{\sum_1^n (y_i - \theta_2)^2/2\right\} \\ &= \exp\{n\theta_2(\theta_2/2 - \overline{y})\} \end{aligned} \tag{53}$$

It is seen that if $\overline{y} = \theta_2/2, K_{12} = 1$ while if $\overline{y} < \theta_2/2, K_{12} > 1$, a result favouring H_1 and if $\overline{y} > \theta_2/2, K_{12} < 1$, evidence against H_1. If $\overline{y} = 0, K_{12} = \exp\{n\theta_2^2/2\} > 1$ with K_{12} larger than larger are n, the sample size and θ_2. Similarly, if $\overline{y} = \theta_2, K_{12} = \exp\{-n\theta_2^2/2\} < 1$ with K_{12} smaller the larger are n and θ_2 Note also that $\partial \log K_{12}/\partial \overline{y} = -n\theta_2$

< 0 indicating that K_{12} that K_{12} is a monotonically decreasing function of $\bar{y}$ with $\theta_2 > 0$

Above two simple hypotheses have been analysed. Posterior odds can also be computed for composite hypotheses which do not involve assigning specific values for all parameters. For example in terms of Example 6, it is possible to compute posterior odds for the two hypotheses. $H_1 : \theta = 0, H_2 : \theta > 0$ or for pairs of the following three hypotheses, $H_1 : \theta = 0$, $H_2 : \theta > 0$ and $H_3 : \theta < 0$. In these cases, there is a simple hypothesis, $\theta = 0$ and composite hypotheses, $\theta \neq 0, \theta > 0$ and $\theta < 0$ In the former case the posterior odds is given by $K_{12} = \Pi_1/\Pi_2 \times \mathrm{BF}_{12}$ with the BF given as follows

$$\mathrm{BF}_{12} = p(\mathbf{y}|\theta = 0) / \int_{-\infty}^{\infty} p(\mathbf{y}|\theta)\pi(\theta)\mathrm{d}\theta \qquad (54)$$

where $\pi(\theta)$ is a prior pdf for θ under $H_2 : \theta \neq 0$. Since in this case the two hypotheses are exhaustive, $K_{12} = P/(1-P)$, where P is the posterior probability for H_1 and $1 - P$ is the posterior probability for H_2. For the three hypotheses, $H_1 : \theta = 0, H_2 : \theta > 0$, and $H_3 : \theta < 0$ posterior odds, K_{12}, K_{13} and K_{23} can be computed given that prior probabilities Π_1, Π_2 and Π_3 for the hypotheses and prior pdfs for θ under H_2 and H_3, $\pi_2(\theta), 0 < \theta < \infty, \pi_3(\theta), -\infty < \theta < 0$ are available. For example, the posterior odds for H_1 versus H_3 and for H_2 and H_3 are:

$$K_{13} = (\Pi_1/\Pi_3)p(\mathbf{y}|\theta = 0) / \int_{-\infty}^{0} p(\mathbf{y}|\theta)\pi_3(\theta)\mathrm{d}\theta \qquad (55)$$

$$K_{23} = (\Pi_2/\Pi_3) \int_{0}^{\infty} p(\mathbf{y}|\theta)\pi_2(\theta)\mathrm{d}\theta \int_{-\infty}^{0} p(\mathbf{y}|\theta)\pi_3(\theta)\mathrm{d}\theta \qquad (56)$$

See Jeffreys (1967), Leamer (1978) and Zellner (1971, 1984) for further analysis of these testing problems. Also treated in these works are regression testing problems, for example computation of posterior odds for the hypotheses $H_1 : \boldsymbol{\beta} = 0$ and $\mathrm{H}_2 : \boldsymbol{\beta} \neq \mathbf{0}$, where β is a vector of regression parameters in the usual linear regression model, $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$. Further, hypotheses referring to sub-vectors of β are considered in these works as well as non-nested regression models. For testing problems in multivariate regression, see Rossi (1980) and Smith and Spiegelhalter (1980). Also, asymptotic approximations to general posterior odds expressions have been considered by Jeffreys (1967), Lindley (1964), Schwarz (1978), Leamer (1978), and Zellner and Rossi (1984).

Finally, it is the case that posterior probabilities associated with alternative hypotheses can be used to obtain estimates and predictions that reflect uncertainties associated with alternative hypotheses. For example, consider the hypotheses for a regression coefficient vector $\boldsymbol{\beta}, H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1$ as given value, and $\boldsymbol{\beta} \neq \boldsymbol{\beta}_1$ with posterior probabilities P and $1 - P$, respectively, computed from the posterior odds $K_{12} = P/(1-P)$. Using a quadratic loss function, $(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' Q (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$ where Q is a given pds matrix and $\tilde{\beta}$ is an estimate, Zellner and vandaele (1975) show that the value of $\tilde{\boldsymbol{\beta}}$ that minimizes expected loss is given by

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^* &= P\boldsymbol{\beta}_1 + (1-P)\overline{\boldsymbol{\beta}}_2 \\ &= \boldsymbol{\beta}_1 + 1/(1+K_{12})(\overline{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_1) \end{aligned} \qquad (57)$$

where $\overline{\boldsymbol{\beta}}_2$ is the posterior mean of β under H_2. It is seen from (56) that the optimal estimate is a simple average of $\overline{\boldsymbol{\beta}}_1$ and $\overline{\boldsymbol{\beta}}_2$ and that from the second line of (56) the estimate can be viewed as a 'shrinkage' estimate with shrinkage factor $1/(1+K_{12})$ It is also possible to perform this analysis when more than two hypotheses are considered.

Bayesian posterior odds have also been derived and used for analysing non-nested models, say two different distributions for a set of observations or two completely different economic models, say a Keynesian model versus a monetarist model or a translog production model versus a Fourier series production model. For details, see Dyer (1973), Geisel (1975), Rossi (1980) and Zellner (1971).

In summary, Bayesian procedures for analysing many different kinds of hypotheses are

available. They involve a statement of uncertainty about alternative hypotheses in the form of prior probabilities and prior distributions for parameters whose values are not specified by the hypotheses under consideration. Using these prior probabilities, likelihood functions, and Bayes' Theorem, posterior odds and probabilities can be computed, analytically or numerically. The posterior odds so obtained provide a representation of views regarding alternative hypotheses that reflects the information in the data.

The Bayesian approach to analysing hypotheses differs markedly from non-Bayesian approaches. In the latter, one hypothesis, the so-called null hypothesis is assumed to be true. A test statistic is chosen, say a *t*-statistic and its distribution under the null hypothesis, assumed to be true is derived. Then the value of the test statistic is computed from the data and compared with what is expected under the null hypothesis. If an unusually large value is obtained, the null hypothesis is rejected. The logic of this procedure seems to parallel that of deductive logic in which a proposition is assumed to be true and then a logical contradiction is deduced which implies that the proposition cannot be true. While this approach is valid in deductive logic, it is not valid in inductive logic wherein all propositions or hypotheses are uncertain and is the reason that Bayesians associate probabilities with hypotheses. Further, as Jaynes (1984) points out, if the null hypothesis is rejected in a non-Bayesian analysis, then so too is the distribution of the test statistic that led to the decision rule for rejection. The fundamental difficulty with the non-Bayesian procedure is that it involves two contradictory assumptions, namely, the null hypothesis is true (with probability one) and the null hypothesis may not be true. In the Bayesian approach, the null hypothesis is not assumed to be true but rather it is assigned a probability between zero and one, a formal representation of an investigator's opinion about the inductive (not deductive) validity of the null hypothesis. As the following quotation from Lehmann (1959) indicates, non-Bayesians frequently have to use such subjective beliefs informally in order to get sensible results:

> Another consideration that frequently enters into the specification of a significance level is the attitude toward the hypothesis before the experiment is performed. If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low (p. 62).

Thus subjective beliefs are frequently employed in non-Bayesian tests but they are not formally incorporated in the theory of such tests in contrast to Bayesian theory in which they are. Further discussion of the comparative features of Bayesian and non-Bayesian testing procedures appears in Jeffreys (1967), Zellner (1971, 1984), Leamer (1978) and Berger (1985).

Robustness Issues in Bayesian Inference

It is desirable that Bayesian inferences and decisions not be overly sensitive to minor departures from assumptions about the forms of (a) prior distributions, (b) likelihood functions, and (c) loss functions. Various procedures have been suggested that attempt to deal with this issue which are called robust procedures. Robust procedures provide users of them some protection from the effects of various possible departures from assumptions but there is a price to be paid for such protection in terms of the precision of inferences. To illustrate, if there is some uncertainty about whether data follow a normal distribution, it is possible to consider a class of distributions containing the normal as a special case, say a Student-*t* distribution. Since the Student-*t* distribution contains more free parameters than a normal distribution, more parameters have to be estimated. If the data distribution is actually normal, there will be a loss of precision in using the Student-*t* distribution. However, if the normal distribution is inappropriate, use of the Student-*t* distribution may produce better results. Further, if there is little information regarding the form of a parametric data distribution, it is possible to use non-parametric methods; see, e.g. Jeffreys (1967, p. 211ff), Ferguson (1967) and Boos and Monahan (1983). In this

last reference, 'boot-strapped' likelihood functions are employed in Bayesian analyses of data.

As regards prior distributions' forms, hierarchical prior distributions are often employed to guard against the possibility of assigning incorrect values to prior distributions' parameters. That is, if $p(\boldsymbol{\theta}|\mathbf{a})$ is a prior pdf for θ and $\mathbf{a}$ is a vector of prior parameters and there is uncertainty about the value of $\mathbf{a}$, the prior pdf can be elaborated as follows, $p(\boldsymbol{\theta}|b)f(\mathbf{a}|\mathbf{b})$ where $f(\mathbf{a}|\mathbf{b})$ is a prior pdf for a with parameter vector $\mathbf{b}$. Then the marginal prior pdf for θ is $p(\boldsymbol{\theta}|\mathbf{b}) = \int p(\boldsymbol{\theta}|\mathbf{a})f(\mathbf{a}|\mathbf{b})\mathrm{d}\mathbf{a}$, an average of $p(\boldsymbol{\theta}|\mathbf{a})$ using $f(\mathbf{a}|\mathbf{b})$ as a weight function. Such hierarchical priors provide some protection against assigning an incorrect value for $\mathbf{a}$ in $p(\boldsymbol{\theta}|\mathbf{a})$ but at a price of increased complexity of analysis. Also, if several prior pdfs are under consideration, say $p_1(\boldsymbol{\theta}|\mathbf{a}_1), p_2(\boldsymbol{\theta}, \mathbf{a}_2), \ldots, p_m(\boldsymbol{\theta}|\mathbf{a}_m)$ it is possible to compute posterior odds given a likelihood function, $l(\theta|\mathbf{y})$ as follows

$$K_{ij} = \left(\Pi_i/\Pi_j\right)\frac{\int l(\boldsymbol{\theta}|\mathbf{y})p_i(\boldsymbol{\theta}|\mathbf{a}_i)\mathrm{d}\theta}{\int l(\boldsymbol{\theta}|\mathbf{y})p_j(\boldsymbol{\theta}|\mathbf{a}_j)\mathrm{d}\boldsymbol{\theta}} \tag{58}$$

where Π_i and Π_j are the prior probabilities associated with prior pdfs i and j, respectively. Assuming that $\sum_{i=1}^{m}\Pi_i = 1$, posterior probabilities $P_1, P_2, \ldots, P_m$, with $\sum_{i=1}^{m}P_i = 1$, can be computed. These posterior probabilities can be employed to average results across different priors, as shown explicitly in (56), and thus to have some protection against using the 'wrong' prior. Also, Berger (1984) suggests using a particular class of prior distributions and checking to determine that inferences are not sensitive to the choice of prior pdf in the particular class. See Kadane (1984) for further discussion of these issues and for suggested measures of robustness.

Last, point estimates and other inference results are often sensitive to the form of the loss function employed. Thus it is important that the effects of possible errors in formulating loss functions be appraised; see Zellner and Geisel (1968), Varian (1975), and Zellner (1984, 1986) for some results relating to this problem area.

Concluding Remarks

An overview of Bayesian inference has indicated that Bayesian inference techniques are available for analysing many basic problems in science. These techniques are noteworthy for their conceptual simplicity and ability to combine prior and sample information in the solution of many scientific inference problems in a coherent manner. Perhaps the most commonly expressed criticism of the Bayesian approach is that it is 'subjective', the implication being that non-Bayesian procedures are 'objective'. In this connection, it is the case that non-Bayesians employ non-sample, subjective information informally in their analyses in choosing significance levels, functional forms for relations, appraising inference results, etc. The eminent non-Bayesian statistician Freedman (1986) has written:

When drawing inferences from data, even the most hardbitten objectivist usually has to introduce assumptions and use prior information. The serious question is how to integrate that information into the inferential process and how to test the assumptions underlying the analysis (p. 127).

In a similar vein, the famous non-Bayesian statistician Tukey (1978) expressed the following views:

> It is my impression that rather generally, not just in econometrics, it is considered decent to use judgement in choosing a functional form, but indecent to use judgment in choosing a coefficient. If judgment about important things is quite all right, why should it not be used for less important ones as well? Perhaps the real purpose of Bayesian techniques is to let us do the indecent thing while modestly concealed behind a formal apparatus. If so, this would not be a precedent. When Fisher introduced the formalities of the analysis of variance in the early 1920s, its most important function was to conceal the fact that the data was being adjusted for block means, an important step forward which if openly visible would have been considered by too many wiseacres of the time to be "cooking the data." If so, let us hope that day will soon come when the role of decent concealment can be freely admitted.... The coefficient may be better estimated from one source or another, or, even best, estimated by economic judgment...

It seems to me a breach of the statistician's trust not to use judgment when that appears to be better than using data (p. 52).

Thus, as is obvious, scientists use both sample and prior information in making inferences. Bayesian inference techniques provide a means of combining these two types of information. In many problems Bayesian inference techniques provide good solutions. Whether Bayesian inference techniques are superior to other inference techniques is an issue that is the subject of much past and current research.

See Also

- ▶ Specification Problems in Econometrics
- ▶ Statistical Decision Theory
- ▶ Statistical Inference
- ▶ Subjective Probability

Bibliography

Aitchison, J., and I.R. Dunsmore. 1975. *Statistical prediction analysis*. Cambridge: Cambridge University Press.

Bayes, T. 1763. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53: 370–418.

Berger, J.O. 1984. The robust Bayesian viewpoint. In *Robustness of Bayesian analysis*, ed. J. Kadane, 64–144. Amsterdam: North-Holland (with discussion).

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*, 2nd ed. New York: Springer.

Blackwell, D., and M.A. Girshick. 1954. *Theory of games and statistical decisions*. New York: Wiley.

Blyth, C.R. 1951. On minimax statistical decision procedures and their admissibility. *Annals of Mathematical Statistics* 22: 22–42.

Boos, D.D., and J.F. Monahan. 1983. Posterior distributions for boot-strapped likelihoods. Unpublished manuscript, North Carolina State University, December.

Box, G.E.P., and G.C. Tiao. 1973. *Bayesian inference in statistical analysis*. Reading: Addison-Wesley.

Boyer, M., and R.E. Kihlstrom (eds.). 1984. *Bayesian models in economic theory*. Amsterdam: North-Holland.

Broemeling, L.D. 1985. *Bayesian analysis of linear models*. New York: Marcel Dekker.

Cox, R.T. 1961. *The algebra of probable inference*. Baltimore: Johns Hopkins University Press.

de Finetti, B. 1970. *The theory of probability*, vol. 2. English trans., New York: Wiley, 1974.

De Groot, M.H. 1970. *Optimal statistical decisions*. New York: McGraw-Hill.

Dyer, A.R. 1973. Discrimination procedures for separate families of hypotheses. *Journal of the American Statistical Association* 68: 970–974.

Edgeworth, F.Y. 1928. In *Contributions to mathematical statistics*, ed. A.L. Bowley. London: Royal Statistical Society. Reprinted, Clifton: Augustus M. Kelley, 1972.

Ferguson, T.S. 1967. *Mathematical statistics: A decision theoretic approach*. New York: Academic Press.

Freedman, D.A. 1986. Reply. *Journal of Business and Economic Statistics* 126–127, January.

Geisel, M.S. 1975. Bayesian comparisons of simple macroeconomic models. In *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, ed. S.E. Fienberg and A. Zellner, 227–256. Amsterdam: North-Holland.

Geisser, S. 1984. On prior distributions for binary trials. *The American Statistician* 38: 244–247.

Good, I.J. 1950. *Probability and the weighing of evidence*. London: Griffin.

Good, I.J. 1965. *The estimation of probabilities*. Cambridge, MA: MIT Press.

Hartigan, J. 1983. *Bayes theory*. New York: Springer.

Heyde, C.C., and I.M. Johnstone. 1979. On asymptotic posterior normality for stochastic processes. *Journal of the Royal Statistical Society, Series B* 41: 184–189.

Jaynes, E.T. 1984. The intuitive inadequacy of classical statistics. *Epistemologia* 7, Special Issue on Probability, Statistics and Inductive Logic, 43–74.

Jeffreys, H. 1967. *Theory of probability*, 3rd ed. Oxford: Clarendon Press (1st ed, 1939).

Jeffreys, H. 1973. *Scientific inference*, 3rd ed. Cambridge: Cambridge University Press. 1st ed, 1931.

Kadane, J.B. (ed.). 1984. *Robustness of Bayesian analysis*. Amsterdam: North-Holland.

Kruskal, W.H. 1978. Tests of significance. In *International encyclopedia of statistics*, vol. 2, ed. W.H. Kruskal and J.M. Tanur, 944–958. New York: The Free Press.

Laplace, P.S. 1820. *Essai philosophique sur les probabilités. English translation as A philosophical essay on probabilities*. New York: Dover. 1951.

Leamer, E.E. 1978. *Specification searches*. New York: Wiley.

Lehmann, E. 1959. *Testing statistical hypotheses*. New York: Wiley.

Lindley, D.V. 1964. The use of prior probability distributions in statistical inference and decisions. In *Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability*, vol. I, ed. J. Neyman, 453–468. Berkeley: University of California Press.

Lindley, D.V. 1965. *Introduction to probability and statistics from a Bayesian viewpoint*, 2 vols. Cambridge: Cambridge University Press.

Lindley, D.V. 1971. *Bayesian statistics: A review*. Philadelphia: Society for Industrial and Applied Mathematics.

Molina, E.C. 1940. Some comments on Bayes' essay. In *Facsimiles of two papers by Bayes*, ed. W.E. Deming, vii–xii. Washington, DC: Graduate School, US Department of Agriculture.

Monahan, J. 1983. Fully Bayesian analysis of ARMA time series models. *Journal of Econometrics* 21: 307–331.

Raiffa, H., and R. Schlaifer. 1961. *Applied statistical decision theory*. Boston: Graduate School of Business Administration, Harvard University.
Rényi, A. 1970. *Foundations of probability*. San Francisco: Holden-Day.
Richard, J.F. 1973. *Posterior and predictive densities for simultaneous equation models*. Berlin: Springer.
Rossi, P.E. 1980. Testing hypotheses in multivariate regression: Bayes vs. non-Bayes procedures. H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago. Paper presented at Econometric Society Meeting, September 1980, Denver.
Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.
Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
Smith, A.F.M., and D.J. Spiegelhalter. 1980. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B* 42: 213–220.
Stigler, S.M. 1983. Who discovered Bayes's Theorem. *The American Statistician* 37: 290–296.
Tiao, G.C., and G.E.P. Box. 1975. Some comments on 'Bayes' estimators. In *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, ed. S.E. Fienberg and A. Zellner, 619–626. Amsterdam: North-Holland.
Tukey, J.W. 1978. Discussion of Granger on seasonality. In *Seasonal analysis of economic time series*, ed. A. Zellner, 50–53. Washington, DC: US Government Printing Office.
Varian, H.R. 1975. A Bayesian approach to real estate assessment. In *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, ed. S.E. Fienberg and A. Zellner, 195–208. Amsterdam: North-Holland.
Wald, A. 1950. *Statistical decision functions*. New York: Wiley.
Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.
Zellner, A. 1977. Maximal data information prior distributions. In *New developments in the applications of Bayesian methods*, ed. A. Aykac and C. Brumat, 211–232. Amsterdam: North-Holland.
Zellner, A. 1984. *Basic issues in econometrics*. Chicago: University of Chicago Press.
Zellner, A. 1986. Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* 81: 446–451.
Zellner, A., and M.S. Geisel. 1968. Sensitivity of control to uncertainty and form of the criterion function. In *The future of statistics*, ed. D.G. Watts, 269–289. New York: Academic Press.
Zellner, A., and P.E. Rossi. 1984. Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics* 25: 365–393.
Zellner, A., and W. Vandaele. 1975. Bayes–Stein estimators for k-means, regression and simultaneous equation models. In *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, ed. S.E. Fienberg and A. Zellner, 317–343. Amsterdam: North-Holland.

Bayesian Methods in Macroeconometrics

Frank Schorfheide

Abstract

This article discusses how Bayesian methods can be used to cope with challenges that arise in the econometric analysis of dynamic stochastic general equilibrium models and vector autoregressions.

Keywords

Bayes' theorem; Bayesian methods in macroeconometrics; Business cycles; Calibration; Cowles Commission; Dynamic macroeconomics; Dynamic stochastic general equilibrium (DSGE) models; Economic growth; Estimation; Expectations; Identification; Statistical inference; Intertemporal optimization problems; Joint probability distributions; Likelihood functions; Macroeconometrics; Misspecification; Monetary policy shocks; Neoclassical growth model; Probability; Rational expectations; Structural change; System-of-equations models; Technology shocks; Vector autoregressions; Linear models; Markov chain Monte Carlo methods; Stochastic growth models; Habit formation; Maximum likelihood; Posterior probability; Regime-switching models; Latent state variables; State-space models

JEL Classifications

D4; D10

Macroeconometrics encompasses a large variety of probability models for macroeconomic time series as well as estimation and inference procedures to study the determinants of economic growth, to examine the sources of business cycle fluctuations, to understand the propagation of shocks, to generate forecasts, and to predict the effects of economic policy changes. Bayesian

methods are a collection of inference procedures that permit researchers to combine initial information about models and their parameters with sample information in a logically coherent manner by use of Bayes' theorem. Both prior and post-data information is represented by probability distributions.

Unfortunately, the term 'macroeconometrics' is often narrowly associated with large-scale system-of-equations models in the Cowles Commission tradition that were developed from the 1950s to the 1970s. These models came under attack on academic grounds in the mid-1970s. Lucas (1976) argued that the models are unreliable tools for policy analysis because they are unable to predict the effects of policy regime changes on the expectation formation of economic agents in a coherent manner. Sims (1980) criticized the fact that many of the restrictions that are used to identify behavioural equations in these models are inconsistent with dynamic macroeconomic theories and proposed the use of vector autoregressions (VARs) as an alternative. Academic research on econometric models in the Cowles tradition reached a trough in the early 1980s and never recovered. The state-of-the-art is summarized in a monograph by Fair (1994).

I am adopting a modern view of macroeconometrics in this article and will portray an active research area that is tied to modern dynamic macroeconomic theory. Reviewing Bayesian methods in macroeconometrics in a short essay is a difficult task. My review is selective and not representative of Bayesian time-series analysis in general. I have chosen some topics that I believe are important, but the list is by no means exhaustive. I focus on the question how Bayesian methods are used to address some of the challenges that arise in the econometric analysis of dynamic stochastic general equilibrium (DSGE) models and VARs. A more extensive treatment can be found in the survey article by An and Schorfheide (2007).

DSGE Models

The term 'DSGE model' is often used to refer to a broad class of dynamic macroeconomic models that spans the standard neoclassical growth model discussed in King et al. (1988) as well as the monetary model with numerous real and nominal frictions developed by Christiano et al. (2005).

A common feature of these models is that decision rules of economic agents are derived from assumptions about preferences and technologies by solving intertemporal optimization problems. Moreover, agents potentially face uncertainty with respect to, for instance, total factor productivity or the nominal interest rate set by a central bank. This uncertainty is generated by exogenous stochastic processes or shocks that shift technology or generate unanticipated deviations from a central bank's interest-rate feedback rule. Conditional on distributional assumptions for the exogenous shocks, the DSGE model generates a joint probability distribution for the endogenous model variables such as output, consumption, investment, and inflation.

What Are the Goals?

While macroeconometric methods are used to address many different questions, several issues stand out. Business cycle analysts are interested in identifying the sources of fluctuations; for instance, how important are monetary policy shocks for movements in aggregate output? We would like to understand the propagation of shocks; for example, what happens to aggregate hours worked in response to a technology shock? Moreover, researchers ask questions about structural changes in the economy: has monetary policy changed in the early 1980s? Why did the volatility of many macroeconomic time series drop in the mid-1980s? Macroeconometricians are also interested in forecasting the future: how will inflation and output growth rates evolve over the next eight quarters? Finally, an important aspect of macroeconometrics is to predict the effect of policy changes: how will output and inflation respond to an unanticipated change in the nominal interest rate? Is it desirable to adopt an inflation targeting regime?

What Are the Challenges?

In principle one could proceed as follows: specify a DSGE model that is sufficiently rich to address

the substantive economic question of interest; derive its likelihood function and fit the model to historical data; answer the questions based on the estimated DSGE model. Unfortunately, this is easier said than done. A trade-off between theoretical coherence and empirical fit poses the first challenge to macroeconometric analysis.

Under certain regularity conditions DSGE models can be well approximated by VARs that satisfy particular cross-coefficient restrictions. The DSGE model is misspecified if these restrictions are at odds with the data and the model has difficulties in tracking and forecasting historical time series. Misspecification was quite apparent for the first generation of DSGE models and has led Kydland, Prescott, and their followers since the early 1980s to abandon formal econometric procedures and advocate a calibration approach, outlined for instance in Kydland and Prescott (1996). Recent Bayesian and non-Bayesian research, however, has resulted in formal econometric tools that are general enough to explicitly account for misspecification problems that arise in the context of DSGE models. Examples of Bayesian approaches are Canova (1994), DeJong et al. (1996), Geweke (1999), Schorfheide (2000), Del Negro and Schorfheide (2004), and Del Negro et al. (2006).

The presence of misspecification might suggest that we should simply ignore the cross-coefficient restrictions implied by dynamic economic theories in the empirical work and try to answer the questions posed above directly by VARs. Unfortunately, there is no free lunch. VARs have many free parameters, and without restrictions on their coefficients tend to generate poor forecasts. VARs do not provide a tight economic interpretation of economic dynamics in terms of the behaviour of rational, optimizing agents. Moreover, it is difficult to predict the effects of rare policy regime changes on the expectation formation and the behaviour of economic agents since these are not explicitly modelled. While the most recent generation of DSGE models comes much closer to matching the empirical fit of VARs, as documented in Smets and Wouters (2003), a trade-off between theoretical coherence and empirical fit remains.

A second challenge is identification. The parameters of a model are identifiable if no two parameterizations of that model generate the same probability distribution for the observables. In VARs the mapping between the one-step-ahead forecast errors of the endogenous variables and the underlying structural shocks is not unique, and additional restrictions are necessary to identify, say, a monetary policy or a technology shock. Many of the popular identification schemes and the controversies surrounding them are surveyed in Cochrane (1994), Christiano and Eichenbaum (1999) and Stock and Watson (2001).

DSGE models can be locally approximated by linear rational expectations (LRE) models. While tightly parameterized compared to VARs, LRE models can generate delicate identification problems. Suppose a model implies that $y_t = \theta \mathrm{E_t}[y_{t-1}] + u_\mathrm{t}$, where u_t is an independently distributed random variable with mean zero. If $0 \leq \theta < 1$, then the only stable law of motion for y_t that satisfies the rational expectations restrictions is $y_t = u_\mathrm{t}$, which means that θ is not identifiable. More elaborate examples are discussed in Beyer and Farmer (2004), Lubik and Schorfheide (2004, 2006), and Canova and Sala (2006). Unfortunately, it is in many cases difficult to detect identification problems in DSGE models, since the mapping from the structural parameters into the autoregressive law of motion for y_t is highly nonlinear and typically can be evaluated only numerically.

Many regularities of macroeconomic time series are indicative of nonlinearities, for instance, the rise and fall of inflation in the 1970s and early 1980s and time- varying volatility of many macroeconomic time series; see, for example, Cogley and Sargent (2005), Sargent et al. (2006), and Sims and Zha (2006). In VARs nonlinear dynamics are typically generated with time-varying coefficients, whereas most DSGE models are nonlinear and only for convenience approximated by linear rational expectations models. Conceptually the analysis of nonlinear models is very similar to the analysis of linear models, but the implementation of the computations is often more cumbersome and poses a third challenge.

How Can Bayesian Analysis Help?

Bayesian analysis is conceptually straightforward. Pre-sample information about parameters is summarized by a prior distribution $p(\theta)$. We can also assign discrete probabilities to distinct models although the distinction between models and parameters is somewhat artificial. The prior is combined with the conditional distribution of the data given the parameters (likelihood function) $p(Y|\theta)$. The application of Bayes' theorem yields the posterior model probabilities and parameter distributions $p(\theta|Y)$. Markov chain Monte Carlo methods can be used to generate θ draws from the posterior. Based on these draws one can numerically approximate the relevant moments of the posterior and make inference about taste and technology parameters as well as the relative importance and the propagation of the various shocks.

The literature on Bayesian estimation of DSGE models began with work by Landon-Lane (1998), DeJong et al. (2000), Schorfheide (2000), and Otrok (2001). DeJong et al. (2000) estimate a stochastic growth model and examine its forecasting performance, Otrok (2001) fits a real business cycle with habit formation and time-to-build to the data to assess the welfare costs of business cycles, and Schorfheide (2000) considers cash-in-advance monetary DSGE models. The Bayesian analysis of VAR dates at least back to Doan et al. (1984).

Since DSGE models are to some extent microfounded, macroeconomists require their parameterization to be consistent with microeconometric evidence on, for instance, labour supply elasticities and the frequency with which firms adjust their prices. If information in the estimation sample were abundant and model misspecification were not a concern, then there would be little need for a prior distribution that summarizes information contained in other data-sets. However, in the estimation of DSGE model this additional information plays an important role.

The prior is used to down-weigh the likelihood function in regions of the parameter space that are inconsistent with out-of-sample information and in which the structural model becomes uninterpretable. The shift from prior to posterior can be an indicator of tensions between different sources of information. If the likelihood function peaks at a value that is at odds with, say, the micro-level information that has been used to construct the prior distribution then marginal data density $\int p(Y|\theta)p(\theta)d\theta$ will be low. If two models have equal prior probabilities, then the ratio of their marginal data densities determine the posterior model odds. Hence, in a posterior odds comparison a DSGE model will automatically be penalized for not being able to reconcile two sources of information with a single set of parameters.

Identification problems manifest themselves through ridges and multiple peaks of equal height in the likelihood function. While Bayesian inference is based on the same likelihood function as classical maximum likelihood estimation, it can bring to bear additional information that may help to discriminate between different parameterizations of a model. If, for instance, the likelihood function is invariant to a subvector θ_1 of θ then the posterior distribution of θ_1 conditional on the remaining parameters will simply equal to the prior distribution. Hence, a comparison of priors and posteriors can provide important insights about the extent to which the data provide information about the parameters of interest. Regardless, the posterior provides a coherent summary of pre-sample and sample information and can be used for inference and decision making. This insight has been used, for instance, by Lubik and Schorfheide (2004) to assess whether monetary policy in the 1970s was conducted in a way that would allow expectations to be self-fulfilling and cause business cycle fluctuations unrelated to fundamental shocks.

Bayesian inference is well suited for model comparisons. Under a loss function that is zero if the correct model is chosen and 1 otherwise, it is optimal to select the model that has the highest posterior probability. However, in many applications, in particular related to the comparison of two possibly misspecified DSGE models, this zero–1 loss function is not very attractive because it does provide little insight into the dimensions along which the structural models should be

improved. Schorfheide provides a framework for the comparison of two or more potentially misspecified DSGE models. A VAR plays the role of a reference model. If the DSGE models are indeed misspecified the VAR will attain the highest posterior probability and the model comparison is based on the question: given a particular loss function, which DSGE model best mimics the dynamics captured by the VAR?

VARs typically have many more parameters than DSGE models and the role of prior distributions is mainly to reduce the effective dimensionality of this parameter space to avoid over-fitting. More interestingly, if one interprets the DSGE model as a set of restrictions on the VAR, then the DSGE model induces a degenerate prior for the VAR coefficients. If the researcher is concerned about potential misspecification of the DSGE model, a natural approach is to relax the DSGE model restrictions and construct a non-degenerate prior distribution that concentrates most of its mass near the restrictions. This approach was originally proposed by Ingram and Whiteman (1994) and has been further developed by Del Negro and Schorfheide (2004), who provide a framework for the joint estimation of VAR and DSGE model parameters. The framework generates a continuum of intermediate specifications that differ according to the degree by which the restrictions are relaxed. This degree is measured by a hyperparameter and the posterior distribution of the hyperparameter can be interpreted as a measure of fit.

Incorporating model and parameter uncertainty into a decision is straightforward in a Bayesian set-up. Levin et al. (2006), for instance, study the effect of optimal monetary policy under parameter uncertainty in the context of an estimated DSGE model. Let δ denote a decision, such as the choice of a monetary policy rule or a tax rate, and L (δ, θ) be a loss function that is used to evaluate the decision. The optimal choice minimizes the posterior risk $\int L(\delta, \theta)p(\theta|Y)d\theta$. The calculation of the risk is facilitated by Markov chain Monte Carlo methods that enable a numerical evaluation of expected losses. If the parameter θ in the loss function is replaced by a future observation y' and $p(\theta|Y)$ is replaced by the predictive distribution $p(y'|Y)$, the decision-theoretic framework can also be used to generate forecasts from the Bayes model.

Finally, with respect to the analysis of nonlinear models, Bayesian methods are in some instances very helpful. Data-augmentation techniques let researchers efficiently deal with numerical complications that arise in models with latent state variables, such as regime-switching models or VARs with time-varying coefficients as in Cogley and Sargent (2005) and Sims and Zha (2006). On the other hand, the need to compute a likelihood function can create serious obstacles. For instance, the computation of the likelihood function for a DSGE model solved with a nonlinear solution method requires a computational-intensive particle filter as in Fernández-Villaverde and Rubio-Ramírez (2006).

Conclusion

The Bayesian paradigm provides a rich framework for inference and decision making with modern macroeconometric models such as DSGE models and VARs. The econometric methods can be tailored to cope with the challenges in this literature: potential model misspecification and a trade-off between theoretical coherence and empirical fit, identification problems, and estimation of models with many parameters based on relatively few observations. Advances in Bayesian computations let the researcher efficiently deal with numerical complications that arise in models with latent state variables, such as regime-switching models, or nonlinear state-space models.

See Also

- ▶ Bayesian Econometrics
- ▶ Markov Chain Monte Carlo Methods
- ▶ Vector Autoregressions

Bibliography

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

Beyer, A., and R. Farmer. 2004. On the indeterminacy of new-Keynesian economics. Working Paper No. 323, European Central Bank.

Canova, F. 1994. Statistical inference in calibrated models. *Journal of Applied Econometrics* 9: S123–SS44.

Canova, F., and L. Sala. 2006. Back to square one: Identification issues in DSGE models. Working Paper No. 583, European Central Bank.

Christiano, L., and M. Eichenbaum. 1999. Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics,* vol. 1A, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113: 1–45.

Cochrane, J. 1994. Shocks. *Carnegie Rochester Conference Series* 41: 295–364.

Cogley, T., and T. Sargent. 2005. Drifts and volatilities: Monetary policies and outcomes in the post WWII U.-S. *Review of Economic Dynamics* 8: 262–302.

DeJong, D., B. Ingram, and C. Whiteman. 1996. A Bayesian approach to calibration. *Journal of Business Economics and Statistics* 14: 1–9.

DeJong, D., B. Ingram, and C. Whiteman. 2000. A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* 98: 201–223.

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

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters. 2006, forthcoming. On the fit of new Keynesian models. *Journal of Business and Economic Statistics.*

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

Fair, R. 1994. *Testing macroeconomic models*. Cambridge, MA: Harvard University Press.

Fernández-Villaverde, J., and Rubio-Ramírez, J. 2006, forthcoming. Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies*.

Geweke, J. 1999. Computational experiments and reality. *Computing in Economics and Finance*, No. 401. Society for Computational Economics, Department of Economics, Boston College.

Ingram, B., and C. Whiteman. 1994. Supplanting the Minnesota prior – Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics* 34: 497–510.

King, R., C. Plosser, and S. Rebelo. 1988. Production, growth, and business cycles: I. Neoclassical model. *Journal of Monetary Economics* 81: 819–840.

Kydland, F., and E. Prescott. 1996. The computational experiment: An econometric tool. *Journal of Economic Perspectives* 10(1): 69–85.

Landon-Lane, J. 1998. Bayesian comparison of dynamic macroeconomic models. PhD thesis, University of Minnesota.

Levin, A., A. Onatski, J. Williams, and N. Williams. 2006. Monetary policy under uncertainty in micro-founded macroeconometric models. In *NBER macroeconomics annual 2005*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

Lubik, T., and F. Schorfheide. 2004. Testing for indeterminacy: An application to U.S. monetary policy. *American Economic Review* 94: 190–217.

Lubik, T., and F. Schorfheide. 2006. A Bayesian look at new open economy macroeconomics. In *NBER macroeconomics annual 2005*, ed. M. Gertler and K. Rogoff. Cambridge, MA: MIT Press.

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

Otrok, C. 2001. On measuring the welfare cost of business cycles. *Journal of Monetary Economics* 47: 61–92.

Sargent, T., N. Williams, and T. Zha. 2006. Shocks and government beliefs: The rise and fall of American inflation. *American Economic Review* 96: 1193–1224.

Schorfheide, F. 2000. Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics* 15: 645–670.

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

Sims, C., and T. Zha. 2006. Were there regime switches in U.S. monetary policy? *American Economic Review* 96: 54–81.

Smets, F., and R. Wouters. 2003. An estimated stochastic dynamic general equilibrium model of the Euro area. *Journal of the European Economic Association* 1: 1123–1175.

Stock, J., and M. Watson. 2001. Vector autoregressions. *Journal of Economic Perspectives* 15(4): 101–116.

Bayesian Non-parametrics

Stephen Graham Walker

Abstract

This article discusses Bayesian nonparametric models, arguing that all Bayesians are constructing probability distributions (the prior) on spaces of density functions. The

parametric Bayesian can be seen to be making restrictive assumptions about the choice of density for modelling data. In contrast, the nonparametric Bayesian constructs a probability distribution on as many densities as possible. The model is infinite dimensional, yet inference is possible, including density estimation and the implementation of decision rules, such as the maximization of expected utility. An example of a nonparametric model is given and a means by which to make inference provided by simulation techniques.

Keywords

Bayesian nonparametrics; Density functions; Expected utility; Latent variables; Likelihood; Markov chain Monte Carlo methods; Parametric models; Probability distribution; Statistical inference; Uncertainty

JEL Classifications

C11; C14

Bayesian nonparametrics, and more generally the Bayesian approach to statistical inference, finds a theoretical justification via a set of axioms of rational behaviour in the presence of uncertainty. Bayesian decision theory establishes how decisions must be made if one desires to avoid irrational behaviour. Thus, coherence is a fundamental concept and is often used as the main argument against competing statistical approaches, such as those based on sampling or fiducial methods. See Lindley (1978) and Bernardo and Smith (1994, ch. 2), who provide a comprehensive discussion on the axiomatic approach to Bayesian inference.

Bayesian statistics is now commonplace among statistical procedures, and is routinely employed in many areas of science, including economics, medicine, biology and others. The use of a prior distribution is the distinguishing feature; the prior distribution updates to the posterior distribution when the data are observed. The prior distribution is assumed to represent subjective beliefs about an unknown parameter; the data then provide further information about the parameter, and the revised beliefs are then to be found in the posterior distribution. The updating mechanism from prior to posterior is formalized through the procedure of multiplying the likelihood function by the prior density function. This idea was apparently first written down by Thomas Bayes in the 18th century.

The uncertainty which frustrates the choice of decision is to be assessed via the use of a probability distribution, and the coherent way to make progress with the inclusion of data is via Bayes's theorem. To elaborate, suppose θ is a parameter to be investigated, which if known would provide a decision, and that θ belongs to the parameter space Θ, which is a finite dimensional space. For example, Θ could represent the real line. Data arise from the density $f(x;\theta)$ in the form of independent and identically distributed observations, say $x_1, \ldots, x_n$, that is, a sample of size n. The likelihood function is any function of θ which is proportional to

$$l(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Let $\pi(\theta)$ denote the prior density function. Then the posterior density function is given by

$$\pi(\theta|\, x_1,\ldots,x_n) = \frac{l(\theta)\pi(\theta)}{\int_\Theta l(\theta)\pi(\mathrm{d}\theta)}.$$

Inference about θ is then performed using the posterior distribution. For example, an estimate of θ could be the posterior mean, which is

$$\hat{\theta} = \int_\Theta \theta\pi(\mathrm{d}\theta|\, x_1,\ldots,x_n).$$

Alternatively, interest might be in the estimate of the density function $f(x;\theta)$ itself. In the Bayesian approach this would be provided by the predictive density function, which is given by

$$\hat{f}(x|\, x_1,\ldots,x_n) = \int_\Theta f(x;\theta)\pi(\mathrm{d}\theta|\, x_1,\ldots,x_n).$$

Making decisions under uncertainty can be undertaken via the maximization of expected utility approach, see Hirshleifer and Riley (1992),

which for the Bayesian would amount to maximizing, over the decision space,

$$\overline{u}(d) = \int_{\Theta} u(d,\theta)\pi(d\theta|\, x_1,...,x_n),$$

where $u(d, \theta)$ is the utility (reward) of selecting decision d from a set of possible decisions when the true parameter state is θ.

The key to the understanding of Bayesian nonparametrics is to think about the family of densities from which the data arose, which in the parametric case is represented as $f(x; \theta)$. Such a family may be known, or assumed to be known, for the data $\{x_1,..., x_n\}$ and the family can be represented by a finite dimensional parameter θ. On the other hand, it may not be known, making assumptions about the family of densities problematic. In this case what is actually unknown is the density function which generated the data: not a parameter, but the entire density function itself. As a Bayesian it is incumbent on the experimenter to construct a prior distribution on the unknown, which is the entire density, and so a probability distribution, the prior, must be placed on the space of density functions. Let such a space be denoted by **F**, so a prior distribution Π must be constructed on **F**.

In fact, any parametric Bayesian model defines a probability measure on **F**. With a parametric model indexed by $\theta \in \Theta$, with family of densities $f(x; \theta)$ and prior $\pi(\theta)$, yields

$$P(f \in A) = \int_{\{\theta: f(\cdot;\theta) \in A\}} \pi(\mathrm{d}\theta),$$

for suitable sets of densities $A \subset \mathbf{F}$. If we let $\Pi(A) = P(f \in A)$, then Π is the prior distribution on F, and the pair $\{f(x; \theta), \pi(\theta)\}$ are a useful way to construct a probability on F. However, this approach of using the parametric model restricts the choice, the A's for which $\Pi(A) > 0$ form a very small set, and so, while it can be seen that all Bayesians are constructing probability measures on **F**, it is the parametric Bayesian who is making restrictive assumptions.

A consequence of the restrictive choice can be seen by considering $\Omega \subset \mathbf{F}$, which we define to be the smallest set of densities which are allocated probability 1, that is, $\Pi(\Omega) = 1$. A parametric family is typically checked off with the data once it has been observed, to see if the model and the data are compatible. Yet this practice is clearly in contradiction (that is, incoherent) with the allocation of probability 1. See Lindsey (1999) for more on this aspect of Bayesian inference. It is the responsibility of the Bayesian to select Ω large enough to make any such checks redundant. This may mean having Ω to be the set of all densities, or at least having the set of A's for which $\Pi(A) > 0$ to be as large as can be achieved.

In the Bayesian nonparametric approach, the prior distribution is placed on **F** directly and there is no finite-dimensional parameter characterizing the random density functions chosen from the prior. The model is infinite-dimensional. The prior is now written as $\Pi(\mathrm{d}f)$ to reflect the fact that there is no parametric θ generating the density f. The likelihood function simply becomes

$$l(f) = \prod_{i=1}^{n} f(x_i)$$

and so the posterior is given by

$$\Pi(\mathrm{d}f|\, x_1,...,x_n) = \frac{\Pi_{i=1}^{n} f(x_i)\Pi(\mathrm{d}f)}{\int_{\mathbf{F}} \Pi_{i=1}^{n} f(x_i)\Pi(\mathrm{d}f)}.$$

Now, for example, the estimate of the density generating the data can be the predictive density, which is

$$\hat{f}(x|\, x_1,...,x_n) = \int_{\mathbf{F}} f(x)\Pi(\mathrm{d}f|\, x_1,...,x_n).$$

For decision theory, if $u(d, f)$ is the utility of decision d when f is the true density, that is, the true density function generating observations, then the maximization of the expected utility rule yields the decision d maximizing

$$\overline{u}(d) = \int_{\mathbf{F}} u(d,f)\Pi(\mathrm{d}f|\, x_1,...,x_n).$$

So what has happened is that we have replaced the finite-dimensional θ with the infinite-dimensional f.

Obviously, the important feature in Bayesian nonparametrics is to be able construct a probability distribution Π on $\mathbf{F}$ such that Ω is large. Suppose $\mathbf{F}$ is the space of density functions defined on the real line. Then, for example, we could choose Π by restricting attention to the normal family of density functions. That is, a random normal density function chosen from Π has the mean μ chosen from the probability density $\pi(\mu)$ and the variance σ^2 chosen from the probability density $\pi(\sigma^2)$.

However, the shape of densities constructed this way is restricted to the normal shape and Ω will not be large. To generate more shapes of density function, one needs to increase the number of parameters from two to a large number, even an infinite, but countable, number. This can be achieved by a mixture model, taking the normal distribution and mixing it over the parameters by using a random distribution function. If we let $\theta = (\mu, \sigma^2)$ and let N denote the normal density function, then a random density function can be obtained via

$$f_P(x) = \int_\Theta N(x|\,\theta)\mathrm{d}P(\theta),$$

where P is a random distribution function defined on $(-\infty, +\infty) \times (0, +\infty)$. The variety of shapes for f_P as P varies over distribution functions is enlarged significantly.

The choice for the random distribution function P needs to be discussed. A common choice is the Dirichlet process model, introduced by Ferguson (1973). The model generates random distribution functions which are discrete. Essentially, a random path (stochastic process) is generated which behaves as a distribution function. That is, it starts at zero and moves to 1 in a non-decreasing way. It is possible to sample a Dirichlet process via the strategy of taking $\{\theta_i\}_{i=1}^{\infty}$ to be independent and identically distributed from some fixed distribution P_0 and $\{v_i\}_{i=1}^{\infty}$ to be independent and identically distributed from beta $(1,c)$ for some $c > 0$. Then

$$P = \sum_{j=1}^{\infty} w_j\, \delta_\theta j,$$

where $w_1 = v_1$ and for $j > 1$,

$$w_j = v_j \prod_{l=1}^{j-1} (1 - v_l\,).$$

It is straightforward to show that the sum of the w_j's is one. It is that $E(P) = P_0$ and for suitable sets B,

$$\mathrm{Var}\{P(B)\} = \frac{P_0(B)\{1 - P_0(B)\}}{c+1}.$$

Using the Dirichlet process itself for modelling independent and identically distributed observations, say $\{y_1, \ldots, y_n\}$, can be done and the posterior is also a Dirichlet process with updated parameters $c \to c + n$ and

$$P_0 \;\to \frac{cP_0 + nP_n}{c+n},$$

where P_n is the empirical distribution function of $\{y_1, \ldots, y_n\}$. Hence, the Bayes estimate is a nice mixture of the prior choice and the empirical distribution.

However, the Dirichlet process is better placed to construct random density functions via mixtures, and we can write the random density function based on the mixture as

$$f_{w;\theta}(x) = \sum_{j=1}^{\infty} w_j N\big(x|\,\theta_j\big).$$

This is an infinite-dimensional model and is known as the mixture of Dirichlet process model. It was first studied by Lo (1984) and can really be estimated only by using recent advances in posterior simulation techniques based on Gibbs samplers and more generally Markov chain Monte Carlo methods (Smith and Roberts 1993; Tierney 1994). The original simulation technique was introduced by Escobar (1988), and since then a number of algorithms have been described. A nice approach, as is becoming usual with Bayesian nonparametric models, is to use latent variables. A slice variable can work well with the mixture of Dirichlet process model by introducing the latent variable u, which has joint density with x given by

$$f_{w;\theta}(x,u) = \sum_{j=1}^{\infty} 1\big(u < w_j\big) N\big(x|\,\theta_j\big).$$

Integrating over u returns us to the original model, and the usefulness of the latent variable is apparent in that it makes the infinite sum finite. That is, there is only a finite number of the $\{w_j\}$ which are greater then u, for each $u > 0$. A Gibbs sampler can now be employed on the model exactly. Typically one is interested in prediction, and at each iteration of the Markov chain it is possible to sample from the predictive density.

There is nowadays a wide range of Bayesian nonparametric models from which to select for any kind of statistical context. See, for example, Walker et al. (1999) for details. Analysis, in the way of inference or decision making, is then typically undertaken using simulation techniques such as Markov chain Monte Carlo methods.

Most Bayesian nonparametric priors are based on stochastic processes. The probability measure for the process acts as the prior distribution. One such example employed in survival models is based on independent increment processes; one has

$$S(t) = e^{-Z(t)},$$

where, with probability 1, Z is a non-decreasing process with $Z(0) = 0$ and $\lim_{t \to +\infty} Z(t) = +\infty$. Here S is a random survival distribution, the law governing the path is the prior. The posterior is also based on an independent increment process (conjugate), and a limiting version of the Bayes estimate turns out to be the Kaplan–Meier nonparametric estimator for a survival function.

Bayesian nonparametric models support more outcomes than parametric models. Prior distributions are constructed on function spaces, such as density functions, survival distribution functions or even hazard functions. The prior distributions are the laws governing stochastic processes whose sample paths behave like these types of functions. Inference is typically reliant on Markov chain Monte Carlo methods, often following the introduction of latent variables.

See Also

- ▶ Bayesian Statistics
- ▶ Decision Theory in Econometrics
- ▶ Utility

Bibliography

Bernardo, J.M., and A.F.M. Smith. 1994. *Bayesian theory*. London: Wiley.

Escobar, M.D. 1988. *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means*. Ph.D. thesis, Department of Statistics, Yale University.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1: 209–230.

Hirshleifer, J., and J.G. Riley. 1992. *The analysis of uncertainty and information*. Cambridge: Cambridge University Press.

Lindley, D.V. 1978. The Bayesian approach (with discussion). *Scandinavian Journal of Statistics* 5: 1–26.

Lindsey, J.K. 1999. Some statistical heresies. *The Statistician* 48: 1–40.

Lo, A.Y. 1984. On a class of Bayesian nonparametric estimates I Density estimates. *Annals of Statistics* 12: 351–357.

Smith, A.F.M., and G.O. Roberts. 1993. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 55: 3–23.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *Annals of Statistics* 22: 1701–1722.

Walker, S.G., P. Damien, P.W. Laud, and A.F.M. Smith. 1999. Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* 61: 485–527.

Bayesian Statistics

José M. Bernardo

Abstract

Statistics is primarily concerned with analysing data, either to assist in appreciating some underlying mechanism or to reach effective decisions. All uncertainties should be described by probabilities, since probability is the only appropriate language for a logic that deals with all degrees of uncertainty, not just absolute truth and falsity. This is the essence of Bayesian statistics. Decision-making is embraced by introducing a utility function and then maximizing expected utility. Bayesian statistics is designed to handle all situations

where uncertainty is found. Since some uncertainty is present in most aspects of life, Bayesian statistics arguably should be universally appreciated and used.

Keywords

Asymptotic behaviour; Bayes, T.; Bayesian reference criterion; Bayesian statistics; Exchangeability; Expected utility; Hypothesis testing; Improper prior function; Inference; Likelihood; Nonparametric models; Nuisance parameters; Point estimation; Prediction; Probability; Reference analysis; Region estimation; Representation theorems; Robustness; Statistical decision theory; Statistical inference; Subjective probability; Sufficiency; Sure thing principle; Uncertainty

JEL Classifications

C11

Bayesian statistics is a comprehensive approach to both statistical inference and decision analysis which derives from the fact that, for rational behaviour, all uncertainties in a problem must necessarily be described by probability distributions.

Unlike most other branches of mathematics, conventional methods of statistical inference do not have an axiomatic basis; as a consequence, their proposed desiderata are often mutually incompatible, and the analysis of the same data may well lead to incompatible results when different, apparently intuitive, procedures are tried. In marked contrast, the Bayesian approach to statistical inference is firmly based on axiomatic foundations which provide a unifying logical structure and guarantee the mutual consistency of the methods proposed. Bayesian methods constitute a complete paradigm for statistical inference, a scientific revolution in Kuhn's sense. Bayesian statistics require only the mathematics of probability theory and the interpretation of probability which most closely corresponds to the standard use of this word in everyday language: a conditional measure of uncertainty. The main consequence of these axiomatic foundations is precisely the requirement to describe with probability distributions all uncertainties present in the problem. Hence, parameters are treated as random variables; this is not a description of their variability (parameters are typically fixed unknown quantities) but a description of the uncertainty about their true values.

The Bayesian paradigm is easily summarized. Thus, if available data D are assumed to have been generated from a probability distribution $p(D|\boldsymbol{\omega})$ characterized by an unknown parameter vector $\boldsymbol{\omega}$, the uncertainty about the value of $\boldsymbol{\omega}$ before the data have been observed must be described by a prior probability distribution $p(\boldsymbol{\omega})$. After data D have been observed, the uncertainty about the value of $\boldsymbol{\omega}$ is described by its posterior distribution $p(\boldsymbol{\omega}|D)$, which is obtained via Bayes's theorem; hence the adjective Bayesian for this form of inference. Point and region estimates for $\boldsymbol{\omega}$ may be derived from $p(\boldsymbol{\omega}|D)$ as useful summaries of its contents. Measures of the compatibility of the posterior with a particular set $\boldsymbol{\Theta}_0$ of parameter values may be used to test the hypothesis $H_0 = \{q \in \boldsymbol{\Theta}_0\}$. If data consist of a random sample $D = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ from a probability distribution $p(\boldsymbol{x}|\boldsymbol{\omega})$, inferences about the value of a future observation $\boldsymbol{x}$ from the same process are derived from the (posterior) predictive distribution $p(\boldsymbol{x}|\ D) = \int_\Omega p(\boldsymbol{x}|\ \boldsymbol{\omega})p(\boldsymbol{\omega}|\ D)d\boldsymbol{\omega}$.

An important particular case arises when either no relevant prior information is readily available, or that information is subjective and an 'objective' analysis is desired, one that is exclusively based on accepted model assumptions and well-documented data. This is addressed by reference analysis which uses information- theoretic concepts to derive the appropriate reference posterior distribution $\pi(\boldsymbol{\omega}|D)$, defined to encapsulate inferential conclusions about the value of $\boldsymbol{\omega}$ solely based on the assumed probability model $p(D|\boldsymbol{\omega})$ and the observed data D.

Pioneering textbooks on Bayesian statistics were Jeffreys (1961), Lindley (1965), Zellner (1971) and Box and Tiao (1973). For modern elementary introductions, see Berry (1996) and Lee (2004). Intermediate to advanced monographs on Bayesian statistics include Berger (1985), Bernardo and Smith (1994), Gelman et al. (2003), O'Hagan (2004) and Robert

(2001). This article may be regarded as a very short summary of the material contained in the forthcoming second edition of Bernardo and Smith (1994). For a recent review of objective Bayesian statistics, see Bernardo (2005) and references therein.

Foundations

The central element of the Bayesian paradigm is the use of probabilities to describe all relevant uncertainties, interpreting $\Pr(A|H)$, the probability of A given H, as a conditional measure of uncertainty, on a [0,1] scale, about the occurrence of the event A in conditions H. There are two different independent arguments which prove the mathematical inevitability of the use of probabilities to describe uncertainties.

Exchangeability and Representation Theorems

Available data often consist of a finite set $\{\mathbf{x}_1, \ldots, \boldsymbol{x}_n\}$ of 'homogeneous' observations, in the sense that only their values matter, not the order in which they appear. Formally, this is captured by the notion of exchangeability. The set of random vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \boldsymbol{x}_j \in \mathscr{X}$, is exchangeable if their joint distribution is invariant under permutations. An infinite sequence of random vectors is exchangeable if all its finite subsequences are exchangeable. Notice that, in particular, any random sample from any model is exchangeable. The general representation theorem implies that, if a set of observations is assumed to be a subset of an exchangeable sequence, then it constitutes a random sample from a probability model $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, described in terms of some parameter vector w; furthermore, this parameter $\boldsymbol{\omega}$ is defined as the limit (as $n \to \infty$) of some function of the observations, and available information about the value of $\boldsymbol{\omega}$ must necessarily be described by some probability distribution $p(\boldsymbol{\omega})$. This formulation includes 'nonparametric' (distribution free) modelling, where $\boldsymbol{\omega}$ may index, for instance, all continuous probability distributions on $\mathscr{X}$. Notice that $p(\boldsymbol{\omega})$ does not model a possible variability of $\boldsymbol{\omega}$ (since $\boldsymbol{\omega}$ will typically be a fixed unknown vector), but models the uncertainty associated with its actual value. Under exchangeability (and therefore under any assumption of random sampling), the general representation theorem provides an existence theorem for a probability distribution $p(\boldsymbol{\omega})$ on the parameter space $\boldsymbol{\Omega}$, and this is an argument which depends only on mathematical probability theory.

Statistical Inference and Decision Theory

Statistical decision theory provides a precise methodology to deal with decision problems under uncertainty, but it also provides a powerful axiomatic basis for the Bayesian approach to statistical inference. A decision problem exists whenever there are two or more possible courses of action. Let $\mathscr{A}$ be the class of possible actions, let $\boldsymbol{\Theta}$ be the set of relevant events which may affect the result of choosing an action, and let $c(a,q) \in \mathscr{C}$, be the consequence of having chosen action a when event $\boldsymbol{\theta}$ takes place. The triplet $\{\mathscr{A}, \boldsymbol{\Theta}, \mathscr{C}\}$ describes the structure of the decision problem. Different sets of principles have been proposed to capture a minimum collection of logical rules that could sensibly be required for rational decision-making. These all consist of axioms with a strong intuitive appeal; examples include the *transitivity* of preferences (if $a_1 \succ a_2$ and $a_2 \succ a_3$, then $a_1 \succ a_3$), and the *sure thing principle* (if $a_1 \succ a_2$ given E, and $a_1 \succ a_2$ given $\overline{E}$, then $a_1 \succ a_2$). Notice that these rules are not intended as a description of actual human decision-making, but as a normative set of principles to be followed by someone who aspires to achieve coherent decisionmaking. There are naturally different options for the set of acceptable principles, but they all lead to the same basic conclusions:

- Preferences among possible consequences should be measured with a *utility* function $u(c) = u(a,q)$ which specifies, on some numerical scale, their desirability.
- The uncertainty about the relevant events should be measured with a probability distribution $p(q|D)$ describing their plausibility given the conditions under which the decision must be taken (assumptions made and available data D).

- The best strategy is to take that action a^* with maximizes the corresponding expected utility, $\int_{\Theta} u(a, q)p(q|\ D)dq$.

Notice that the argument described above establishes (from another perspective) the need to quantify the uncertainty about all relevant unknown quantities (the actual value of the vector $\boldsymbol{\theta}$), and specifies that this must have the mathematical structure of a probability distribution. It has been argued that the development described above (which is not qsted when decisions have to be made) does not apply to problems of statistical inference, where no specific decision making is envisaged. Notice, however, that (*a*) a problem of statistical inference is typically considered worth analysing because it may eventually help make sensible decisions (as Ramsey put it in the 1930s, a lump of arsenic is poisonous because it may kill someone, not because it has actually killed someone), and (b) statistical inference on $\boldsymbol{\theta}$ has the mathematical structure of a decision problem, where the class of alternatives is the functional space of all possible conditional probability distributions of $\boldsymbol{\theta}$ given the data, and the utility function is a measure of the amount of information about $\boldsymbol{\theta}$ which the data may be expected to provide.

In statistical inference it is often convenient to work in terms of the nonnegative loss function $\ell(a, q) = \sup_{a \in \mathscr{A}}\{u(a, q)\} - u(a, q)$, which directly measures, as a function of $\boldsymbol{\theta}$, the *penalty* for choosing a wrong action. The undesirability of each possible action $a \in \mathscr{A}$ is then measured by its *expected loss,* $l(a|\ D) = \int_{\Theta} \ell(a, q)p(q|\ D)dq$ and the best action a^* is that with the minimum expected loss.

The Bayesian Paradigm

The statistical analysis of some observed data set $D \in \mathscr{D}$ typically begins with some informal descriptive evaluation, which is used to suggest a tentative, formal probability model $\{p(D|\omega, \mathrm{H}), \omega \in \boldsymbol{\Omega}\}$ which, given some assumptions H, is supposed to represent, for some (unknown) value of $\boldsymbol{\omega}$, the probabilistic mechanism which has generated the observed data D. The arguments outlined above establish the logical need to assess a prior probability distribution $p(\boldsymbol{\omega}|H)$ over the parameter space $\boldsymbol{\Omega}$, describing the available knowledge about the value of $\boldsymbol{\omega}$ under the accepted assumptions H, prior to the data being observed. It then follows from Bayes's theorem that, if the probability model is correct, all available information about the value of $\boldsymbol{\omega}$ after the data D have been observed is contained in the corresponding *posterior distribution,*

$$p(\boldsymbol{\omega}|\, D, H) = \frac{p(D|\ \boldsymbol{\omega}, H)p(\boldsymbol{\omega}|\, H)}{\int_{\Omega} p(D|\ \boldsymbol{\omega}, H)p(\boldsymbol{\omega}|\, H)d\boldsymbol{\omega}}, \boldsymbol{\omega} \in \boldsymbol{\Omega} \tag{1}$$

It is this systematic use of Bayes's theorem to incorporate the information provided by the data that justifies the adjective 'Bayesian' by which the paradigm is usually known. It is obvious from Bayes's theorem that any value of $\boldsymbol{\omega}$ with zero prior density will have zero posterior density. Thus, it is typically assumed (by appropriate restriction, if necessary, of the parameter space $\boldsymbol{\Omega}$) that prior distributions are strictly positive. To simplify the presentation, the assumptions H are often omitted from the notation, but the fact that all statements about $\boldsymbol{\omega}$ given D are also conditional to H should always be kept in mind.

Computation of posterior densities is often facilitated by noting that Bayes's theorem may be simply expressed as $p(\omega\backslash D) \propto p(D\backslash\boldsymbol{\omega})p(\boldsymbol{\omega})$ (where $\propto$ stands for 'proportional to' and where, for simplicity, the assumptions H have been omitted from the notation), since the missing proportionality constant $\left[\int_{\Omega} p(D|\ \boldsymbol{\omega})p(\boldsymbol{\omega})d\boldsymbol{\omega}\right]^{-1}$ may always be deduced from the fact that $p(\boldsymbol{\omega}\backslash D)$, a probability density, must integrate to 1.

Improper Priors

An improper prior function is defined as non-negative function $\pi(\boldsymbol{\omega})$ such that $\int_{\Omega} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ is not finite. The formal expression of Bayes's theorem remains, however technically valid if $p(\boldsymbol{\omega})$ is replaced by an improper prior function $n(\boldsymbol{\omega})$, provided the proportionality constant exists,

thus leading to a well-defined *proper* posterior density $\pi(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\pi(\boldsymbol{\omega})$, which does integrate to 1.

Likelihood Principle

Considered as a function of $\boldsymbol{\omega}$ for fixed data D, $p(D|\boldsymbol{\omega})$ is often referred to as the likelihood function. Thus, Bayes's theorem is simply expressed in words by the statement that the posterior is proportional to the likelihood times the prior. It follows from (1) that, provided the same prior $p(\boldsymbol{\omega})$ is used, two different data sets D_1 and D_2, with possibly different probability models $p_1(D_1|\boldsymbol{\omega})$ and $p_2(D_2|\boldsymbol{\omega})$ which yield proportional likelihood functions, will produce identical posterior distributions for $\boldsymbol{\omega}$. This immediate consequence of Bayes's theorem has been proposed as a principle on its own, the *likelihood principle,* and it is seen by many as an obvious requirement for reasonable statistical inference. In particular, for any given prior $p(\boldsymbol{\omega})$, the posterior distribution does not depend on the set $\mathscr{D}$ of possible data values (the *outcome space*). Notice, however, that the likelihood principle applies only to inferences about the parameter vector $\boldsymbol{\omega}$ once the data have been obtained. Consideration of the outcome space is essential, for instance, in model criticism, in the design of experiments, in the derivation of predictive distributions, and in the construction of objective Bayesian procedures.

Sequential Learning

Naturally, the terms 'prior' and 'posterior' are only relative to a particular set of data. As one would expect, if exchangeable data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed. Indeed, $p(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1}|\ \boldsymbol{\omega})\ p(\boldsymbol{\omega}\ |\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$, for $i = 1, \ldots, n-1$, so that the 'posterior' at a given stage becomes the 'prior' at the next.

Sufficiency

For a given probability model, one may find that some particular function of the data $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathscr{T}$ is a sufficient statistic in the sense that, given the model, $\boldsymbol{t}(D)$ contains all information about $\boldsymbol{\omega}$ which is available in D. Formally, $\boldsymbol{t}$ is sufficient if (and only if) there exist non-negative functions f and g such that the likelihood function may be factorized in the form $p(D|\boldsymbol{\omega}) = f\ (\boldsymbol{\omega},\ \boldsymbol{t})g(D)$. A sufficient statistic always exists, for $\boldsymbol{t}(D) = D$ is obviously sufficient; however, a much simpler sufficient statistic, with a fixed dimensionality which is independent of the sample size, often exists. In fact this is known to be the case whenever the probability model belongs to the *generalized exponential family,* which includes many of the more frequently used probability models. It is easily established that if $\boldsymbol{t}$ is sufficient, then the posterior distribution of $\boldsymbol{\omega}$ depends only on the data D through $\boldsymbol{t}(D)$, and $p(\boldsymbol{\omega}|D) = p(\boldsymbol{\omega}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\omega})\ p(\boldsymbol{\omega})$.

Robustness

As one would expect, for fixed data and model assumptions, different priors generally lead to different posteriors. Indeed, Bayes' theorem may be described as a data-driven probability transformation machine which maps prior distributions (describing prior knowledge) into posterior distributions (representing combined prior and data knowledge). It is important to analyse the robustness of the posterior to changes in the prior. Objective posterior distributions based on reference priors (see below) play a central role in this context. Investigation of the sensitivity of the posterior to changes in the prior is an important ingredient of the comprehensive analysis of the sensitivity of the final results to all accepted assumptions, which any responsible statistical study should contain.

Nuisance Parameters

Typically, the quantity of interest is not the whole parameter vector $\boldsymbol{\omega}$, but some function $q = q(\boldsymbol{\omega})$ of possibly lower dimension than $\boldsymbol{\omega}$. Any valid conclusion on the value of $\boldsymbol{\theta}$ will be contained in its posterior probability distribution $p(q|D)$, which may be derived from $p(\boldsymbol{\omega}|D)$ by standard use of probability calculus. Indeed, if $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\omega}) \in \Lambda$ is some other function of $\boldsymbol{\omega}$ such that $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}\}$ is a one-to-one transformation of $\boldsymbol{\omega}$, and $J(\boldsymbol{\omega}) = (\partial\psi/\partial\boldsymbol{\omega})$ is the corresponding Jacobian matrix, one may change variables to obtain $p(\psi|D) = p(q, \boldsymbol{\lambda}|D) = p(\boldsymbol{\omega}|D)/|J(\boldsymbol{\omega})|$, and the required

posterior of $\boldsymbol{\theta}$ is $p(q|\ D) = \int_{\Lambda} p(q, \boldsymbol{\lambda}|\ D)d\boldsymbol{\lambda}$, the marginal density obtained by integrating out the nuisance parameter $\boldsymbol{\lambda}$. Naturally, introduction of $\boldsymbol{\lambda}$ is not necessary if $\boldsymbol{\theta}(\boldsymbol{\omega})$ is a one-to-one transformation of $\boldsymbol{\omega}$. Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm, is a difficult (often polemic) problem for conventional statistics.

Restricted Parameter Space

Sometimes, the range of possible values of $\boldsymbol{\omega}$ is effectively restricted by contextual considerations. If $\boldsymbol{\omega}$ is known to belong to $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$, the prior distribution is positive only in $\boldsymbol{\Omega}_c$ and, if one uses Bayes's theorem, it is immediately found that the restricted posterior is

$$p(\boldsymbol{\omega}|\ D, \boldsymbol{\omega} \in \boldsymbol{\Omega}_c) = p(\boldsymbol{\omega}|\ D)/\int_{\boldsymbol{\Omega}_c} p(\boldsymbol{\omega}|\ D)d\boldsymbol{\omega},$$

for $\boldsymbol{\omega} \in \boldsymbol{\Omega}_c$ (and obviously vanishes if $\boldsymbol{\omega} \notin \boldsymbol{\Omega}_c$). Thus, to incorporate a restriction on the possible values of the parameters, it suffices to renormalize the unrestricted posterior distribution to the set $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$ of parameter values which satisfy the required condition. Incorporation of known constraints on the parameter values, a simple renormalization within the Bayesian paradigm, is another very difficult problem for conventional statistics.

Asymptotic Behaviour

The behaviour of posterior distributions when the sample size is large is important, for at least two different reasons: (*a*) asymptotic results provide useful first-order approximations when actual samples are relatively large, and (*b*) objective Bayesian methods typically depend on the asymptotic properties of the assumed model. Let $D = \{\boldsymbol{x}1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_j \in \mathcal{X}$, be a random sample of size n from $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. It may be shown that, as $n \to \infty$, the posterior distribution $p(\boldsymbol{\omega}|D)$ of a discrete parameter $\boldsymbol{\omega}$ typically converges to a degenerate distribution which gives probability one to the true value of $\boldsymbol{\omega}$, and that the posterior distribution of a continuous parameter $\boldsymbol{\omega}$ typically converges to a normal distribution centred at its maximum likelihood estimate (MLE) $\hat{\boldsymbol{\omega}}$, with a covariance matrix $F^{-1}(\hat{\boldsymbol{\omega}})/n$, where $F(\boldsymbol{\omega})$ is Fisher information matrix, of general element

$$F_{ij}(\boldsymbol{\omega}) = -E_{\boldsymbol{x}|\boldsymbol{\omega}}\left[\partial^2 \log[p(\boldsymbol{x}|\ \boldsymbol{\omega})]/\left(\partial\boldsymbol{\omega}_i \partial\boldsymbol{\omega}_j\right)\right].$$

Prediction

When data consist of a set $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of homogeneous observations, one is often interested in predicting the value of a future observation $\boldsymbol{x}$ generated by the same random mechanism that has generated the observations in D. It follows from the foundations arguments discussed above that the solution to this prediction problem must be a probability distribution $p(\boldsymbol{x}|D)$ which describes the uncertainty about the value that $\boldsymbol{x}$ will take, given the information provided by D, and any other available knowledge. In particular, if contextual information suggests that data D may be considered to be a random sample from a distribution in the family $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and $p(\boldsymbol{\omega})$ is a probability distribution which encapsulates all available prior information on the value of $\boldsymbol{\omega}$, the corresponding posterior will be (by Bayes's theorem) $p(\boldsymbol{\omega}|\ D) \propto \prod_{j=1}^{n} p\left(\boldsymbol{x}_j|\ \boldsymbol{\omega}\right)p(\boldsymbol{\omega})$. Since $p(\boldsymbol{x}|\boldsymbol{\omega}, D) = p(\boldsymbol{x}|\boldsymbol{\omega})$, the total probability theorem may then be used to obtain the desired posterior *predictive* distribution

$$p(\boldsymbol{x}|\ D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x}|\ \boldsymbol{\omega})p(\boldsymbol{\omega}|\ D)d\boldsymbol{\omega} \qquad (2)$$

which has the form of a *weighted average:* the average of all possible probability distributions of $\boldsymbol{x}$, weighted with their corresponding posterior densities. Notice that the conventional practice of plugging in some point estimate $\widetilde{\boldsymbol{\omega}} = \widetilde{\boldsymbol{\omega}}(D)$ and using $p(\boldsymbol{x}|\widetilde{\boldsymbol{\omega}})$ to predict $\boldsymbol{x}$ may be seriously misleading, for this totally ignores the uncertainty about the true value of $\boldsymbol{\omega}$. If the assumptions on the probability model are correct, the posterior predictive distribution $p(\boldsymbol{x}|D)$ will converge, as the sample size increases, to the distribution $p(\boldsymbol{x}|D)$ which has generated the data. Indeed, a good technique to assess the quality of the inferences about $\boldsymbol{\omega}$ encapsulated in $p(\boldsymbol{\omega}|D)$ is to check against the observed data the predictive distribution $p(\boldsymbol{x}|D)$ generated from $p(\boldsymbol{\omega}|D)$. The argument used to

derive $p(\boldsymbol{x}|D)$ may be extended to obtain the predictive distribution of any function $\boldsymbol{y}$ of future observations generated by the same process, namely, $p(\boldsymbol{y}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{y}|\ \boldsymbol{\omega}) p(\boldsymbol{\omega}|\ D)$.

Reference Analysis

The posterior distribution combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information is too vague to warrant the effort required to have it formalized in the form of a probability distribution, or it is too subjective to be useful in scientific communication or public decision making. It is therefore important to identify the mathematical form of a reference prior, a prior that would have a minimal effect, relative to the data, on the posterior inference. Much work has been done to formulate priors which would make this idea mathematically precise. This section summarizes an approach, based on information theory, which may be argued to provide the most advanced general procedure available. In this formulation, the reference prior is that which maximizes the missing information about the quantity of interest.

Reference Distributions

Consider data D, generated by a random mechanism $p(D|\theta)$ which depends only on a real-valued parameter $\theta \in \Theta \subset \mathscr{R}$, and let $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathscr{T}$ be any sufficient statistic (which may well be the complete data set D). In Shannon's general information theory, the amount of information $I\{\mathscr{T}, p(\theta)\}$ which may be expected to be provided by D, about the value of θ is

$$I\{\mathscr{T}, p(\theta)\} = \int_{\mathscr{T}} \int_{\Theta} p(\boldsymbol{t}, \theta) \log \frac{p(\boldsymbol{t}, \theta)}{p(\boldsymbol{t}) p(\theta)} d\theta d\boldsymbol{t} = E_t \left[\int_{\Theta} p(\theta|\ \boldsymbol{t})\ \log \frac{p(\theta)\boldsymbol{t}}{p(\theta)} d\theta \right] \tag{3}$$

the expected logarithmic divergence of the prior from the posterior. This is a *functional* of the prior distribution $p(\theta)$: the larger the prior information, the smaller the information which the data may be expected to provide. The functional $I\{\mathscr{T}, p(\theta)\}$ is concave, non-negative, and invariant under one-to-one transformations of θ. Consider now the amount of information $I\{\mathscr{T}^k, p(\theta)\}$ about θ which may be expected from the experiment which consists of k conditionally independent replications $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k\}$ of the original experiment. As $k \to \infty$, such an experiment would provide any *missing information* about θ which could possibly be obtained within this framework; thus, as $k \to \infty$, the functional $I\{\mathscr{T}^k, p(\theta)\}$ will approach the *missing information* about θ associated with the prior $p(\theta)$. Intuitively, the reference prior for θ is that which maximizes the missing information about θ. If $\pi_k(\theta|\ \mathscr{P})$ denotes the prior density which maximizes $I\{\mathscr{T}^k, p(\theta)\}$ in the class $\mathscr{P}$ of strictly positive prior distributions which are compatible with accepted assumptions on the value of θ (which may well be the class of all strictly positive proper priors), then the θ-reference prior $\pi(\theta|\ \mathscr{P})$ is the limit of the sequence of priors $\{\pi_k(\theta|\ \mathscr{P})\}_{k=1}^{\infty}$. The limit is taken in the precise sense that, for any value of the sufficient statistic $\boldsymbol{t}$, the reference posterior, the pointwise limit $\pi(\theta|\ \boldsymbol{t}, \mathscr{P})$ of the corresponding sequence of posteriors $\{\pi_k(\theta|\ \boldsymbol{t}, \mathscr{P})\}_{k=1}^{\infty}$ where $\pi_k(\theta|\boldsymbol{t}, \mathscr{P}) \propto p(\boldsymbol{t}|\theta)\pi_k(\theta|\ \mathscr{P})$, may be obtained from $\pi(\theta|\ \mathscr{P})$ by formal use of Bayes' theorem, so that $\pi(\theta|\boldsymbol{t}, \mathscr{P}) \propto p(\boldsymbol{t}|\theta)\pi(\theta|\ \mathscr{P})$.

The limiting procedure in the definition of a reference prior is not some kind of asymptotic approximation, but an essential element of the definition, required to capture the basic concept of missing information. Notice that, by definition, reference distributions depend only on the asymptotic behaviour of the assumed probability model, a feature which greatly simplifies their actual derivation.

Reference prior *functions* are often simply called reference priors, even though they are usually improper. They should not be considered as expressions of belief, but technical devices to obtain (proper) posterior distributions, which are a limiting form of the posteriors that would have been obtained from prior beliefs which, when compared with the information which data could provide, are relatively uninformative with respect to the quantity of interest.

If θ may take only a *finite* number m of different values, the missing information about θ associated to the prior $p(\theta)$ is its entropy, $H\{p(\theta)\} = -\sum_{j=1}^{m} p(\theta_j)\log p(\theta_j)$. Hence the reference prior $\pi(\theta|\mathscr{P})$ is in this case is the prior with *maximum entropy* within $\mathscr{P}$. In particular, if $\mathscr{P}$ contains all priors over $\{\theta 1,\ldots, \theta_m\}$, then the reference prior when θ is the quantity of interest is the uniform prior $\pi(\theta) = \{1/m, \ldots, 1/m\}$.

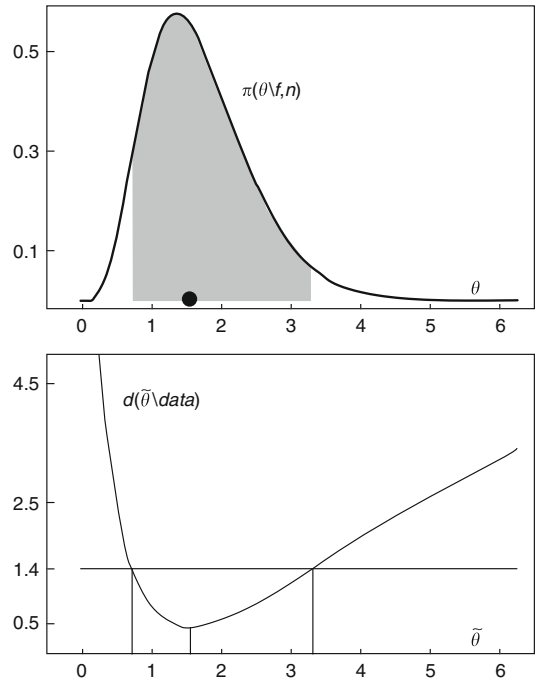
If the sufficient statistic t is a consistent, asymptotically sufficient estimator $\widetilde{\theta}$ of a continuous parameter θ, and the class of priors is the set $\mathscr{P}_0$ of all strictly positive priors, then the reference prior is simply

$$\pi(\theta|\,\mathscr{P}_0) \propto p\left(\theta|\widetilde{\theta}\right)|_{\tilde{\theta}=\theta} \propto p\left(\theta|\widetilde{\theta}\right)|\widetilde{\theta} = \theta, \quad (4)$$

where $p\left(\widetilde{\theta}|\,\theta\right)$ is any asymptotic approximation to the posterior distribution of θ, and $p\left(\widetilde{\theta}|\,\theta\right)$ is the sampling distribution of $\widetilde{\theta}$. Under conditions which guarantee asymptotic posterior normality, this reduces to *Jeffreys prior*, $\pi(\theta_0|\,\mathscr{P}) \propto F(\theta)^{1/2}$, where $F(\theta)$ is Fisher information function. One-parameter reference priors are consistent under re-parametrization; thus, if $\psi = \psi(\theta)$ is a piecewise one-to-one function of θ, then the ψ-reference prior is simply the appropriate probability transformation of the θ-reference prior.

Example 1. Exponential Data If $\boldsymbol{x} = \{x_1,\ldots, x_n\}$ is a random sample from $\theta e^{-e\ x}$, the reference prior is Jeffreys prior $\pi(\theta) = \theta^{-1}$, and the reference posterior is a gamma distribution $\pi(\theta|x) = Ga(\theta|n, t)$, where $t = \sum_{j=1}^{n} x_j$. With a random sample of size $n = 5$ (simulated from an exponential distribution with $\theta = 2$), which yielded a sufficient statistic $t = \sum_j x_j = 2.949$, the result is represented in the upper panel of Fig. 1. Inferences about the value of a future observation from the same process may are described by the reference predictive posterior

$$\pi(x|\boldsymbol{t}) = \int_0^{\infty} \theta e^{-\theta x}\mathrm{Ga}(\theta|\,n,t)d\theta = n\ t^n(x+t)^{-(n+1)}.$$



Bayesian Statistics, Fig. 1 Bayesian reference analysis of the parameter θ of an exponential distribution $p(x|\theta) = \theta\, e^{-x\theta}$, given a sample of size $n = 5$ with $t = \sum_j x_j = 2.949$

Nuisance Parameters

The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the problem to a sequential application of the established procedure for the single parameter case. Thus, if one drops explicit mention to the class $\mathscr{P}$ of priors compatible with accepted assumptions to simplify notation, if the probability model is $\{p(\boldsymbol{t}|\theta, \lambda), \theta \in \Theta, \lambda \text{ e } \Lambda\}$ and a θ-reference prior $\pi_\theta(\theta, \lambda)$ is required, the reference algorithm proceeds in two steps:

1. Conditional on θ, $p(\boldsymbol{t}|\theta, \lambda)$ depends only on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the conditional reference prior $\pi(\lambda|\theta)$.
2. If $\pi(\lambda|\theta)$ is proper, this may be used to integrate out the nuisance parameter, thus obtaining the one-parameter integrated model

$$p(\boldsymbol{t}|\,\theta) = \int_{\Lambda} p(\boldsymbol{t}|\,\theta,\lambda)\ \pi\ (\lambda|\,\theta)\ d\lambda$$

to which the one-parameter algorithm may be applied again to obtain $\pi(\theta)$. The θ-reference prior is then $\pi_\theta(\theta, \lambda) = \pi(\lambda|\theta)\ \pi(\theta)$, and the required reference posterior is $\pi(\theta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\theta)\pi(\theta)$.

If the conditional reference prior $\pi(\lambda|\theta)$ is not proper, then the procedure is performed within an increasing sequence $\{\Lambda_i\}$ of subsets converging to Λ over which $\pi(\lambda|\theta)$ is integrable. This makes it possible to obtain a corresponding sequence of θ-reference posteriors $\{\pi_i(\theta|\ \boldsymbol{t})\}$ for the quantity of interest θ, and the required reference posterior is the corresponding pointwise limit $\pi(\theta|\boldsymbol{t}) = \lim_i \pi_i(\theta|\boldsymbol{t})$.

The θ-reference prior does not depend on the choice of the nuisance parameter λ. Notice, however, that the reference prior may depend on the parameter of interest; thus, the θ-reference prior may differ from the φ-reference prior unless either φ is a piecewise one-to-one transformation of θ or φ is asymptotically independent of θ. This is an expected consequence of the fact that the conditions under which the missing information about θ is maximized may be different from the conditions under which the missing information about some function $\varphi = \varphi(\theta, \lambda)$ is maximized.

The preceding algorithm may be generalized to any number of parameters.

Thus, if the model is $p(\boldsymbol{t}|\omega_1,\ldots, \omega_m)$, a reference prior $\pi(\theta_m|\theta_{m-1;}\ldots, \theta_1) \times \cdot \times \pi(\theta_2|\theta_1) \times \pi(\theta_1)$ may sequentially be obtained for each ordered parametrization $\{\theta_1(\boldsymbol{\omega}),\ldots, \theta_m(\boldsymbol{\omega})\}$ of interest, and these are invariant under re-parametrization of any of the $\theta_i(\omega)$'s. The choice of the ordered parametrization $\{\theta_1,\ldots, \theta_m\}$ precisely describes the particular prior required.

Flat Priors

Mathematical convenience often leads to the use of 'flat' priors, typically some limiting form of a convenient family of priors; this may, however, have devastating consequences. Consider, for instance, that in a normal setting $p(\boldsymbol{x}|\, m) = N_k(\bar{x}, n^{-1}I)$, inferences are desired on $\theta = \sum_{i=1}^k \mu_i^2$, the squared distance of the unknown mean $\boldsymbol{\mu}$ to the origin. It is easily verified that the posterior distribution of θ based on a uniform prior on $\boldsymbol{\mu}$ (or in any 'flat' proper approximation) is strongly inconsistent (Stein's paradox). This is due to the fact that a uniform (or nearly uniform) prior on $\boldsymbol{\mu}$ is highly informative about θ, introducing a severe bias on its marginal posterior. The reference prior which corresponds to a parametrization of the form $\{\theta, \boldsymbol{\lambda}\}$ produces, however, for any choice of the nuisance parameter vector $\boldsymbol{\lambda}$, a reference posterior $\pi(\theta|\,\boldsymbol{x}, \mathscr{P}_0) \propto \theta^{-1/2}\chi^2(nt|\, k, n\theta)$, where $t = \sum_{i=1}^k \bar{x}_i^2$,. Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior distributions derived from multivariate 'flat' priors (proper or improper) is indeed very frequent. Hence, sloppy, uncontrolled use of 'flat' priors (rather than the relevant reference priors) should be very strongly discouraged.

Inference Summaries

From a Bayesian perspective, the final outcome of a problem of inference about any unknown quantity is the corresponding posterior distribution. Thus, given some data D and conditions H, all that can be said about any function $q = q(\boldsymbol{\omega})$ of the parameters which govern the model is contained in the posterior distribution $p(q|D, H)$, and all that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y}|D, H)$. However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to summarize the information contained in the posterior distribution by (a) providing values of the quantity of interest which, in the light of the data, are likely to be a good proxy for its true (unknown) value, and by (b) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. The Bayesian counterparts of those of traditional problems of estimation and hypothesis testing are now briefly considered.

Point Estimation

Let D be the available data, which are assumed to have been generated by a probability model $\{p(D|\boldsymbol{\omega}),\ \omega \in \boldsymbol{\Omega}\}$, and let $q = q(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be

the quantity of interest. A *point estimator* of $\boldsymbol{\theta}$ is some function of the data $\tilde{q} = \tilde{q}(D)$ which could be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$. Formally, to choose a point estimate for $\boldsymbol{\theta}$ is a decision problem, where the action space is the class $\boldsymbol{\Theta}$ of possible $\boldsymbol{\theta}$ values. As dictated by the foundations of decision theory, to solve this decision problem it is necessary to specify a loss function $\ell(\tilde{q}, q)$ measuring the consequences of acting as if the true value of the quantity of interest were $\boldsymbol{\theta}$, when it is actually $\tilde{q}$. The expected posterior loss if $\boldsymbol{\theta}$ were used is

$$l(\tilde{q}\,|\,D) = \int_{\Theta} \ell(\tilde{q}, q) p(q\,|\,D)\; dq, \qquad (5)$$

and the corresponding *Bayes estimator* is that function of the data, $q^* = q^*(D)$, which minimizes $l(\tilde{q}\,|\,D)$.

For any given model, data and prior, the Bayes estimator obviously depends on the loss function which has been chosen. The loss function is context specific, and should be selected in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for scientific communication. These loss functions produce estimates which may often be regarded as simple descriptions of the location of the posterior distribution. If the loss function is quadratic, so that $\ell(\tilde{q}, q) = (\tilde{q} - q)^t (\tilde{q} - q)$, the corresponding Bayes estimator is the posterior mean $E[q|D]$ (on the assumption that the mean exists). Similarly, if the loss function is a zero-one function, so that $\ell(\tilde{q}, q) = 0$ if $\tilde{q}$ belongs to a ball or radius ε centred in $\boldsymbol{\theta}$ and $\ell(\tilde{q}, q) = 1$ otherwise, the corresponding Bayes estimator converges to the posterior mode as the ball radius $\boldsymbol{\varepsilon}$ tends to zero (on the assumption that a unique mode exists). If θ is univariate and the loss function is linear, so that $\ell\left(\widetilde{\theta}, \theta\right) = c_1\left(\widetilde{\theta} - \theta\right)$ if $\widetilde{\theta} \geq \theta$, and $\ell\left(\widetilde{\theta}, \theta\right) = c_2\left(\widetilde{\theta} - \theta\right)$ otherwise, the Bayes estimator is the posterior quantile of order $c_2/(c_1 + c_2)$, so that $Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. In particular, if $c_1 = c_2$, the corresponding Bayes estimator is the posterior median. The results quoted for linear loss functions clearly illustrate the fact that any possible parameter value may turn out be a Bayes estimator: it all depends on the loss function characterizing the consequences of the anticipated uses of the estimate.

Conventional loss functions are typically non-invariant under re-parametrization, so that the Bayes estimator φ^* of a one-to-one transformation $\varphi = \varphi(q)$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\varphi(\boldsymbol{\theta}^*)$ (the univariate posterior median, which is invariant, is an interesting exception). Moreover, conventional loss functions focus on the discrepancy between the estimate $\tilde{q}$ and the true value $\boldsymbol{\theta}$, rather then on the more relevant discrepancy between the probability models which they label. Intrinsic losses directly focus on the discrepancy between the probability distributions $p(D\tilde{q})$ and $\delta(\tilde{q}, q)$, and typically produce invariant solutions. An attractive example is the intrinsic discrepancy $\delta(\tilde{q}, q)$, defined as the minimum logarithmic divergence between a probability model labelled by $\boldsymbol{\theta}$ and a probability model labelled by $\tilde{q}$. When there are no nuisance parameters, this is

$$\begin{aligned}\delta(\tilde{q}, q) &= \min\{\kappa(\tilde{q}\,|\,q),\;\; \kappa(q\,|\,\tilde{q})\}, \kappa\left(q_i\,|\,\tilde{q}_j\right) \\ &= \int_{\mathscr{T}} p\left(\boldsymbol{t}\,|\,q_j\right) \log \frac{p\left(\boldsymbol{t}\,|\,q_j\right)}{p\left(\boldsymbol{t}\,|\,q_i\right)} d\boldsymbol{t}, \end{aligned} \qquad (6)$$

where $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathscr{T}$ is *any* sufficient statistic (which may well be the whole data set D). The definition is easily extended to problems with nuisance parameters. The Bayes estimator is obtained by minimizing the corresponding posterior expected loss. An objective estimator, the *intrinsic estimator* $\tilde{q}_{int} = \tilde{q}_{int}(D)$, is obtained by minimizing the expected intrinsic discrepancy with respect to the *reference* posterior distribution,

$$d(\tilde{q}\,|\,d) = \int_{\boldsymbol{\Theta}} \delta(\tilde{q}, q) \pi(q\,|\,D) dq \qquad (7)$$

Since the intrinsic discrepancy is invariant under re-parametrization, minimizing its posterior expectation produces *invariant* estimators. Thus, the intrinsic estimator of say, the log of the speed of a galaxy is simply log of the intrinsic estimator of the speed of the galaxy.

Region Estimation

To describe the inferential content of the posterior distribution of the quantity of interest $p(q|D)$ it is often convenient to quote *credible* regions, defined as subsets of the parameter space $\boldsymbol{\Theta}$ of given posterior probability. For example, the identification of regions containing 50, 90, 95, or 99 per cent of the probability under the posterior may be sufficient to convey the general quantitative messages implicit in $p(q|D)$. Indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots. A posterior *q-credible region* for $\boldsymbol{\theta}$ is any region $C \subset \boldsymbol{\Theta}$ such that $\int_C p(q|\, D)dq = q$. Notice that this provides immediately a direct intuitive statement about the unknown quantity of interest $\boldsymbol{\theta}$ in probability terms, in marked contrast to the circumlocutory statements provided by conventional confidence intervals. A credible region is invariant under re-parametrization; thus, for any q-credible region C for $\boldsymbol{\theta}$, $\varphi(C)$ is a q-credible region for $\varphi = \varphi(q)$.

Clearly, for any given q there are generally infinitely many credible regions. Credible regions are often selected to have minimum size (length, area, volume), resulting in highest probability density (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are not invariant under re-parametrization: the image $\varphi(C)$ of an HPD region C will be a credible region for φ, but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. In one-dimensional problems, posterior quantiles are often used to derive credible regions. Thus, if $\theta_q = d_q\,(D)$ is the 100q per cent posterior quantile of $\theta \in \Theta \subset \mathscr{R}$, then $C = \{\theta;\, \theta \leq \theta_q\}$ is a onesided, typically unique q-credible region, and it is invariant under re-parametrization; the similarly invariant probability centred q-credible regions of the form $C = \left\{\theta;\, \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\right\}$ are easier to compute than HPD regions; this notion, however, does not extend to multivariate problems.

Choosing a p-credible region may be seen as a decision problem where the action space is the class of all p-credible regions. Foundations then dictate that a loss function $\ell(\tilde{q}, q)$ must be specified, and that the region chosen should consist of those $\widetilde{\boldsymbol{\theta}}$ values with the lowest expected posterior loss $l(\tilde{q}|\, D) = \int_\Theta \ell(\tilde{q}|\, q)p(q|\, D)dq$. By definition, lowest posterior loss (LPL) regions are credible regions where all points in the region have smaller expected posterior loss than all points outside. If the loss function is quadratic, so that $\ell(\tilde{q}, q) = (\tilde{q} - q)^t(\tilde{q} - q)$, the LPL p-credible region is a Euclidean sphere centred at the posterior mean $E[\boldsymbol{\theta}|D]$. Like HPD regions, LDL quadratic credible regions are not invariant under re-parametrization; however, LDL intrinsic regions, which minimize the posterior expectation of the invariant intrinsic discrepancy loss (6) are obviously invariant. *Intrinsic p-credible* regions are LDL intrinsic regions which minimize the expected intrinsic discrepancy with respect to the reference posterior distribution. These provide a general, invariant, objective solution to multivariate region estimation. The notions of point and region parameter estimation described above may easily extended to prediction problems by using the posterior predictive rather than the posterior of the parameter.

Hypothesis Testing

The posterior distribution $p(q\backslash D)$ of the quantity of interest $\boldsymbol{\theta}$ conveys immediate intuitive information on those values of $\boldsymbol{\theta}$ which, given the assumed model, may be taken to be *compatible* with the observed data D, namely, those with a relatively high probability density. Sometimes, a *restriction* $q \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ of the possible values of the quantity of interest (where $\boldsymbol{\Theta}_0$ may possibly consist of a single value $\boldsymbol{\theta}_0$) is suggested in the course of the investigation as deserving special consideration, either because restricting $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}_0$ would greatly simplify the model or because there are additional, context-specific arguments suggesting that $q \in \boldsymbol{\Theta}_0$. Intuitively, the *hypothesis* $H_0 \equiv \{q \in \boldsymbol{\Theta}_0\}$ should be judged to be *compatible* with the observed data D if there are elements in Θ_0 with a relatively high posterior density; however, a more precise conclusion is often required and, once again, this is possible with a decision-oriented approach. Formally, testing the hypothesis $H_0 = \{q \in \boldsymbol{\Theta}_0\}$ is a *decision problem* where the action space has only two elements, namely, to

accept (a_0) or to reject (a_1) the proposed restriction. To solve this decision problem, it is necessary to specify an appropriate loss function, $\ell(a_i, q)$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value $\boldsymbol{\theta}$ of the vector of interest. The optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_{\boldsymbol{\Theta}} \ell(a_0, q) p(q|\, D)\ dq$, is larger than the expected posterior loss of rejecting, $\int_{\boldsymbol{\Theta}} \ell(a_1, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\, D)\ d\boldsymbol{\theta}$, that is, if (and only if)

$$\int_{\boldsymbol{\Theta}} [\ell(a_0, q) - \ell(a_1, q)] p(q|\, D) dq = \int_{\boldsymbol{\Theta}} \Delta\ell(q) p(q|\, D) dq > 0 \tag{8}$$

Therefore, only the loss difference $\Delta\ell(q) = \ell(a_0, q) - \ell(a_1, q)$, which measures the *advantage* of rejecting H_0 as a function of $\boldsymbol{\theta}$, has to be specified: the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive.

The simplest loss structure has the zero-one form given by $\{\ell(a_0, q) = 0,\ \ell(a_1, q) = 1\}$ if $q \in \Theta_0$ and, similarly, $\{\ell(a_0, q) = 1, \ell(a_1, q) = 0\}$ if $q \notin \boldsymbol{\Theta}_0$, so that the *advantage* $\Delta\ell(q)$ of rejecting H_0 is 1 if $q \notin \boldsymbol{\Theta}_0$ and it is – 1 otherwise. With this, rather naive, loss function the optimal action is to reject H_0 if (and only if) $\Pr(q \notin \boldsymbol{\Theta}_0|D) > \Pr(q \notin \boldsymbol{\Theta}_0|D)$. Notice that this formulation requires that $\Pr(q \notin \boldsymbol{\Theta}_0) > 0$, that is, that the hypothesis H_0 has a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter and $\boldsymbol{\Theta}_0$ consists of a single point $\boldsymbol{\theta}_0$ (sharp null problems), this requires the use of a non-regular highly informative prior which places a positive probability mass at $\boldsymbol{\theta}_0$. This posterior probability approach is therefore only appropriate if it is sensible to condition on the assumption that $\boldsymbol{\theta}$ is indeed concentrated around $\boldsymbol{\theta}_0$.

Frequently, however, the compatibility of the observed data with H_0 is to be judged without assuming such a sharp prior knowledge. In those situations, the advantage $\Delta\ell(q)$ of rejecting H_0 as a function of $\boldsymbol{\theta}$ may be typically assumed to be of the general form $\Delta\ell(q) = \delta(\Theta_0, q) - d^*$, for some $d^* > 0$, where $\delta(\boldsymbol{\Theta}_0, q)$ is some measure of the discrepancy between the assumed model $p(D|q)$ and its closest approximation within the class $\{p(D|q_0),\ q_0 \in \boldsymbol{\Theta}_0\}$ and such that $\delta(\Theta_0, q) = 0$ whenever $q\ e\ \boldsymbol{\Theta}_0$, and d^* is a context dependent *utility constant* which measures the (necessarily positive) advantage of being able to work with the restricted model when it is true. For reasons similar to those supporting its use in estimation, an attractive choice for the loss function $\delta(\boldsymbol{\Theta}_0, q)$ is an appropriate extension of the intrinsic discrepancy loss; when there are no nuisance parameters, this is given by $\delta(\boldsymbol{\Theta}_0, q) = \inf_{q_0 \in\ \boldsymbol{\Theta}_0} \delta(q_0, q)$ where $\delta(q_0, q)$ is the intrinsic discrepancy loss defined by (6). The corresponding optimal strategy, called the 'Bayesian reference criterion' (BRC), is then to reject H_0 if, and only if,

$$d(\boldsymbol{\Theta}_0|\, D) = \int_{\Theta} \delta(\boldsymbol{\Theta}_0, q) \pi(q|\, D)\ dq > d^*. \tag{9}$$

The choice of d^* plays a similar role to the choice of the significance level in conventional hypothesis testing. Standard choices for scientific communication may be of the form $d^* = \log k$ for, in view of (6) and of (7), this means that the data D are expected to be at least k times more likely under the true model than under H_0. This is actually equivalent to rejecting H_0 if $\boldsymbol{\Theta}_0$ is not contained in an intrinsic q_k-credible region for $\boldsymbol{\theta}$ whose size q_k depends on k. Under conditions for asymptotic posterior normality,

$$q_k \approx 2\Phi\left[(2\ \log\ k - 1)^{1/2}\right] - 1,$$

where Φ is the standard normal distribution function. For instance, if $k = 100$, $q_k \approx 0.996$, while if $k = 11.25$, $q_k \approx 0.95$. The Bayesian reference criterion provides a general objective procedure for multivariate hypothesis testing which is invariant under re-parametrization.

Example 2. Exponential Data, Continued The intrinsic discrepancy loss for an exponential model is $\delta\left(\widetilde{\theta}, \theta\right) = g(\varphi)$, if $\varphi \leq 1$, and $\delta\left(\widetilde{\theta}, \theta\right) = g(1/\varphi)$, if $\varphi > 1$, where $g(\varphi) = \varphi - 1 - \log \varphi$,

and $\varphi = \widetilde{\theta}/\theta$. Using (7) with the data from Example 1, the expected intrinsic loss $d\left(\widetilde{\theta}|\, x\right)$ is the function represented in the lower panel of Fig. 1. The intrinsic estimate is the value which minimizes $d\left(\widetilde{\theta}|\, x\right)$, $\widetilde{\theta}_{int} = 1.546$ (marked with a solid dot in the figure), and the intrinsic 0.90-credible set is (0.720,3.290), the set of parameter values with expected loss below 1.407 (corresponding to the shaded area in the upper panel of the figure).

See Also

- Bayes, Thomas (1702–1761)
- Bayesian Econometrics
- Bayesian Methods in Macroeconometrics
- Bayesian Non-parametrics
- Bayesian Time Series Analysis
- de Finetti, Bruno (1906–1985)
- Savage, Leonard J. (Jimmie) (1917–1971)
- Statistical Decision Theory

Bibliography

Berger, J. 1985. *Statistical decision theory and bayesian analysis*. New York: Springer.

Bernardo, J. 2005. Reference analysis. In *Handbook of statistics 25*, ed. D. Dey and C. Rao, 17–90. Amsterdam: North-Holland.

Bernardo, J.M., and A.F.M. Smith. 1994. *Bayesian theory*. Chichester: Wiley. 2nd edn available in 2007.

Berry, D. 1996. *Statistics: A bayesian perspective*. Belmont: Wadsworth.

Box, G., and G. Tiao. 1973. *Bayesian inference in statistical analysis*. New York: Wiley Classics.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2003. *Bayesian data analysis*. 2nd ed. London: Chapman and Hall.

Jeffreys, H. 1961. *Theory of probability*. 3rd ed. Oxford: University Press.

Lee, P. 2004. *Bayesian statistics: An introduction*. 3rd ed. London: Arnold.

Lindley, D. 1965. *Introduction to probability and statistics from a bayesian viewpoint*. Cambridge: Cambridge University Press.

O'Hagan, A. 2004. *Bayesian inference*. 2nd ed. London: Arnold.

Robert, C. 2001. *The Bayesian choice*. 2nd ed. New York: Springer.

Zellner, A. 1971. *An introduction to bayesian inference in econometrics*. New York: Wiley.

Bayesian Time Series Analysis

Mark F. J. Steel

B

Abstract

This article describes the use of Bayesian methods in the statistical analysis of time series. The use of Markov chain Monte Carlo methods has made even the more complex time series models amenable to Bayesian analysis. Models discussed in some detail are ARIMA models and their fractionally integrated counterparts, state space models, Markov switching and mixture models, and models allowing for time-varying volatility. A final section reviews some recent approaches to nonparametric Bayesian modelling of time series.

Keywords

ARCH models; ARFIMA models; ARIMA models; ARMA models; Bayes factor; Bayesian inference; Bayesian methods in econometrics; Bayesian model averaging; Bayesian nonparametrics; Bayesian time series analysis; Business cycles; Cointegration; Computational algorithms; Conditional likelihood; Continuous-time models; Convergence clubs; Data augmentation; Dirichlet processes; Forecasting; GARCH models; Gibbs sampler; Growth regressions; Hidden Markov models; Impulse response function; Kalman filter; latent states; Lévy processes; Long-memory models; Macroeconomic forecasting; Markov chain Monte Carlo methods; Markov switching models; Metropolis Hastings sampler; Nonparametric models; Ornstein–Uhlenbeck processes; Posterior odds; Prediction; Prior odds; Regime switching models; Regression; Sequential learning; Spatial statistics; State space models; Stochastic volatility models; Survival analysis; Threshold autoregressive models; Time series analysis; Uncertainty; Unit roots; Vector autoregressions

JEL Classification
C11; C22

Bayesian Methods

The importance of Bayesian methods in econometrics has increased rapidly since the early 1990s. This has, no doubt, been fuelled by an increasing appreciation of the advantages that Bayesian inference entails. In particular, it provides us with a formal way to incorporate the prior information we often possess before seeing the data, it fits perfectly with sequential learning and decision making, and it directly leads to exact small sample results. In addition, the Bayesian paradigm is particularly natural for prediction, since we take into account all parameter or even model uncertainty. The predictive distribution is the sampling distribution where the parameters are integrated out with the posterior distribution and provides exactly what we need for forecasting, often a key goal of time-series analysis.

Usually, the choice of a particular econometric model is not pre-specified by theory, and many competing models can be entertained. Comparing models can be done formally in a Bayesian framework through so-called posterior odds, which is the product of the prior odds and the Bayes factor. The Bayes factor between any two models is the ratio of the likelihoods integrated out with the corresponding prior and summarizes how the data favour one model over another. Given a set of possible models, this immediately leads to posterior model probabilities. Rather than choosing a single model, a natural way to deal with model uncertainty is to use the posterior model probabilities to average out the inference (on observables or parameters) corresponding to each of the separate models. This is called Bayesian model averaging. The latter was already mentioned in Leamer (1978) and recently applied to economic problems in, for example, Fernández et al. (2001) (for growth regressions) and in Garratt et al. (2003) and Jacobson and Karlsson (2004) (for macroeconomic forecasting).

An inevitable prerequisite for using the Bayesian paradigm is the specification of prior distributions for all quantities in the model that are treated as unknown. This has been the source of some debate, a prime example of which is given by the controversy over the choice of prior on the coefficients of simple autoregressive models. The issue of testing for a unit root (deciding whether to difference the series before modelling it through a stationary model) is subject to many difficulties from a sampling-theoretical perspective. Comparing models in terms of posterior odds provides a very natural Bayesian approach to testing, which does not rely on asymptotics or approximations. It is, of course, sensitive to how the competing models are defined (for example, do we contrast the stationary model with a pure unit root model or a model with a root larger than or equal to 1?) and to the choice of prior. The latter issues have lead to some controversy in the literature, and prompted a special issue of the *Journal of Applied Econometrics* with animated discussion around the paper by Phillips (1991). The latter paper advocated the use of Jeffreys' principles to represent prior ignorance about the parameters (see also the discussion in Bauwens et al. 1999, ch. 6).

Like the choice between competing models, forecasting can also be critically influenced by the prior. In fact, prediction is often much more sensitive than parameter inference to the choice of priors (especially on autoregressive coefficients) and Koop et al. (1995) show that imposing stationarity through the prior on the autoregressive coefficient in a simple AR(1) model need not lead to stabilization of the predictive variance as the forecast horizon increases.

Computational Algorithms

Partly, the increased use of Bayesian methods in econometrics is a consequence of the availability of very efficient and flexible algorithms for conducting inference through simulation in combination with ever more powerful computing facilities, which have made the Bayesian analysis of non-standard problems an almost routine activity. Particularly, Markov chain Monte Carlo (MCMC) methods have opened up a very useful class of computational algorithms and have created a

veritable revolution in the implementation of Bayesian methods. Whereas Bayesian inference before 1990 was at best a difficult undertaking in practice, reserved for a small number of specialized researchers and limited to a rather restricted set of models, it has now become a very accessible procedure which can fairly easily be applied to almost any model. The main idea of MCMC methods is that inference about an analytically intractable posterior (often in high dimensions) is conducted through generating a Markov chain which converges to a chain of drawings from the posterior distribution. Of course, predictive inference is also immediately available once one has such a chain of drawings. Various ways of constructing such a Markov chain exist, depending on the structure of the problem. The most commonly used are the Gibbs sampler and the Metropolis Hastings sampler. The use of data augmentation (that is, adding auxiliary variables to the sampler) can facilitate implementation of the MCMC sampler, so that often the analysis is conducted on an augmented space including not only the model parameters but also things like latent variables and missing observations. An accessible reference to MCMC methods is, for example, Gamerman (1997).

As a consequence, we are now able to conduct Bayesian analysis of time series models that have been around for a long time (such as ARMA models) but also of more recent additions to our catalogue of models, such as Markov switching and nonparametric models, and the literature is vast. Therefore, I will have to be selective and will try to highlight a few areas which I think are of particular interest. I hope this can give an idea of the role that Bayesian methods can play in modern time series analysis.

ARIMA and ARFIMA Models

Many models used in practice are of the simple ARIMA type, which have a long history and were formalized in Box and Jenkins (1970). ARIMA stands for 'autoregressive integrated moving average' and an ARIMA(p,d,q) model for an observed series $\{y_t\}$, $t = 1,\ldots, T$ is a model where the dth difference $z_t = y_t - y_{t-d}$ is taken to induce stationarity of the series. The process $\{z_t\}$ is then modelled as $z_t = \mu + \varepsilon_t$ with

$$\varepsilon_t = \varphi_1\varepsilon_{t-1} + \cdots + \varphi_p\varepsilon_{t-p} + u_t - \theta_1 u_{t-1} - \cdots - \theta_q u_{t-q}$$

or in terms of polynomials in the lag operator L (defined through $L^s x_t = x_{t-s}$):

$$\varphi(L)\varepsilon_t = \theta(L)u_t$$

where $\{u_t\}$ is white noise and usually distributed as $u_t \sim N(0, \sigma^2)$. The stationarity and invertibility conditions are simply that the roots of $\phi(L)$ and $\theta(L)$, respectively, are outside the unit circle. An accessible and extensive treatment of the use of Bayesian methods for ARIMA models can be found in Bauwens et al. (1999). The latter book also has a useful discussion of multivariate modelling using vector autoregressive (VAR) models and cointegration.

The MCMC samplers used for inference in these models typically use data augmentation. Marriott et al. (1996) use a direct conditional likelihood evaluation and augment with unobserved data and errors to conduct inference on the parameters (and the augmented vectors $\varepsilon_a = \left(\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{1-p}\right)'$ and $u_a = \left(u_0, u_{-1}, \ldots, u_{1-q}\right)$). A slightly different approach is followed by Chib and Greenberg (1994), who consider a state space representation and use MCMC on the parameters augmented with the initial state vector.

ARIMA models will either display perfect memory (if there are any unit roots) or quite short memory with geometrically decaying autocorrelations (in the case of a stationary ARMA model). ARFIMA ('autoregressive fractionally integrated moving average') models (see Granger and Joyeux 1980) have more flexible memory properties, due to fractional integration which allows for hyperbolic decay.

Consider $z_t = \Delta y_t - \mu$, which is modelled by an ARFIMA(p,δ,q) model as:

$$\varphi(L)(1-L)^{\delta} z_t = \theta(L)u_t,$$

where $\{u_t\}$ is white noise with $u_t \sim N(0, \sigma^2)$, and $\delta \in (-1, 0{:}5)$. The fractional differencing operator $(1 - L)^{\delta}$ is defined as

$$(1-L)^{\delta} = \sum_{j=0}^{\infty} c_j(\delta) L^j,$$

where $c_0(\cdot) = 1$ and for $j > 0$:

$$c_j(a) = \prod_{k=1}^{j} \left(1 - \frac{1+a}{k}\right).$$

This model takes the entire past of z_t into account, and has as a special case the ARIMA(p,1,q) for y_t (for $\delta = 0$). If $\delta > -1$, z_t is invertible (Odaki 1993) and for $\delta < 0.5$ we have stationarity of z_t. Thus, we have three regimes:

- $\delta \in (-1, -0.5)$: y_t trend-stationary with long memory
- $\delta \in (-0.5,0)$: z_t stationary with intermediate memory
- $\delta \in (0,0.5)$: z_t stationary with long memory.

Of particular interest is the impulse response function $I(n)$, which captures the effect of a shock of size one at time t on y_{t+n}, and is given by

$$I(n) = \sum_{i=0}^{n} c_i(-\delta - 1) J(n-i),$$

with $J(i)$ the standard ARMA(p,q) impulse responses (that is, the coefficients of $\phi^{-1}(L)\theta(L)$). Thus, $I(\infty)$ is 0 for $\delta < 0$, $\theta(1)/\phi(1)$ for $\delta = 0$ and ∞ for $\delta > 0$. Koop et al. (1997) analyse the behaviour of the impulse response function for real US GNP data using a set of 32 possible models containing both ARMA and ARFIMA models for z_t. They use Bayesian model averaging to conduct predictive inference and inference on the impulse responses, finding about one-third of the posterior model probability concentrated on the ARFIMA models. Koop et al. (1997) use importance sampling to conduct inference on the parameters, while MCMC methods are used in Pai and Ravishanker (1996) and Hsu and Breidt (2003).

State Space Models

The basic idea of such models is that an observable y_t is generated by an observation or measurement equation

$$y_t = F_t' \theta_t + v_t,$$

where $v_t \sim N(0; V_t)$, and is expressed in terms of an unobservable state vector θ_t (capturing, for example, levels, trends or seasonal effects) which is itself dynamically modelled through a system or transition equation

$$\theta_t = G_t \theta_{t-1} + w_t,$$

with $w_t \sim N(0, W_t)$ and all error terms $\{v_t\}$ and $\{w_t\}$ are mutually independent. Normality is typically assumed, but is not necessary and a prior distribution is required to describe the initial state vector θ_0. Models are defined by the (potentially time-varying) quadruplets $\{F_t,G_t,V_t,W_t\}$ and the time-varying states θ_t make them naturally adaptive to changing circumstances. This feature also fits very naturally with Bayesian methods, which easily allow for sequential updating. These models are quite general and include as special cases, for example, ARMA models, as well as stochastic volatility models, used in finance (see below).

There is a relatively long tradition of state space models in econometrics and a textbook treatment can already be found in Harvey (1981). Bayesian methods for such models were discussed in, for example, Harrison and Stevens (1976), and a very extensive treatment is provided in West and Harrison (1997), using the terminology 'dynamic linear models'. An accessible introduction to Bayesian analysis with these models can be found in Koop (2003, Ch. 8).

Online sequential estimation and forecasting with the simple Normal state space model above can be achieved with Kalman filter recursions, but more sophisticated models (or estimation of some aspects of the model besides the states) usually require numerical methods for inference. In that case, the main challenge is typically the

simulation of the sequence of unknown state vectors. Single-state samplers (updating one state vector at a time) are generally less efficient than multi-state samplers, where all the states are updated jointly in one step. Efficient algorithms for multi-state MCMC sampling schemes have been proposed by Carter and Kohn (1994) and de Jong and Shephard (1995). For fundamentally non-Gaussian models, the methods in Shephard and Pitt (1997) can be used. A recent contribution of Harvey et al. (2006) uses Bayesian methods for state space models with trend and cyclical components, exploiting informative prior notions regarding the length of economic cycles.

Markov Switching and Mixture Models

Markov switching models were introduced by Hamilton (1989) and essentially rely on an unobserved regime indicator s_t, which is assumed to behave as a discrete Markov chain with, say, K different levels. Given $s_t = i$ the observable y_t will be generated by a time series model which corresponds to regime i, where $i = 1, \ldots, K$. These models are often stationary ARMA models, and the switching between regimes will allow for some non-stationarity, given the regime allocations. Such models are generally known as hidden Markov models in the statistical literature.

Bayesian analysis of these models is very natural, as that methodology provides an immediate framework for dealing with the latent states, $\{s_t\}$, and a simple MCMC framework for inference on both the model parameters and the states was proposed in Albert and Chib (1993). A bivariate version of the Hamilton model is analysed in Paap and van Dijk (2003), who also examine the cointegration relations between the series modelled and find evidence for cointegration between US per capita income and consumption. Using a similar model, Smith and Summers (2005) examine the synchronization of business cycles across countries and find strong evidence in favour of the multivariate Markov switching model over a linear VAR model.

When panel data are available, another relevant question is whether one can find clusters of entities (such as countries or regions) which behave similarly, while allowing for differences between the clusters. This issue is addressed from a fully Bayesian perspective in Frühwirth-Schnatter and Kaufmann (2006), where model-based clustering (across countries) is integrated with a Markov switching framework (over time). This is achieved by a finite mixture of Markov switching autoregressive models, where the number of elements in the mixture corresponds to the number of clusters and is treated as an unknown parameter. Frühwirth-Schnatter and Kaufmann (2006) analyse a panel of growth rates of industrial production in 21 countries and distinguish two clusters with different business cycles. This also feeds into the important debate on the existence of so-called convergence clubs in terms of income per capita as discussed in Durlauf and Johnson (1995) and Canova (2004).

Another popular way of inducing nonlinearities in time series models is through so-called threshold autoregressive models, where the choice of regimes is not governed by an underlying Markov chain but depends on previous values of the observables. Bayesian analyses of such models can be found in, for example, Geweke and Terui (1993) and are extensively reviewed in Bauwens, Lubrano and Richard (1999, ch. 8). The use of Bayes factors to choose between various nonlinear models, such as threshold autoregressive and Markov switching models is discussed in Koop and Potter (1999).

Geweke and Keane (2006) present a general framework for Bayesian mixture models where the state probabilities can depend on observed covariates. They investigate increasing the number of components in the mixture, as well as the flexibility of the components and the specification of the mechanism for the state probabilities, and find their mixture model approach compares well with ARCH-type models (as described in the next section) in the context of stock return data.

Models for Time-Varying Volatility

The use of conditional heteroskedasticity initially introduced in the ARCH (autoregressive conditional heteroskedasticity) model of Engle (1982)

has been extremely successful in modelling financial time series, such as stock prices, interest rates and exchange rates. The ARCH model was generalized to GARCH (generalized ARCH) by Bollerslev (1986). A simple version of the GARCH model for an observable series $\{y_t\}$, given its past which is denoted by I_{t-1}, is the following:

$$y_t = u_t \sqrt{h_t} \tag{1}$$

where $\{u_t\}$ is white noise with mean zero and variance one. The conditional variance of y_t given I_{t-1} is then h_t, which is modelled as

$$h_t = \omega + \sum_{i=1}^{p} \alpha_i y_{t-1}^2 + \sum_{j=1}^{q} \beta_i h_{t-i} \tag{2}$$

where all parameters are positive and usually $p = q = 1$ is sufficient in practical applications. Bayesian inference for such models was conducted through importance sampling in Kleibergen and van Dijk (1993) and, with MCMC methods, in Bauwens and Lubrano (1998).

An increasingly popular alternative model allows for the variance h_t to be determined by its own stochastic process. This is the so-called stochastic volatility model, which in its basic form replaces (2) by the assumption that the logarithm of the conditional volatility is driven by its own AR(1) process

$$\ln(h_t) = \alpha + \delta \ln(h_{t-1}) + v_t,$$

where $\{v_t\}$ is a white noise process independent of $\{u_t\}$ in (1). Inference in such models requires dealing with the latent volatilities, which are incidental parameters and have to be integrated out in order to evaluate the likelihood. MCMC sampling of the model parameters and the volatilities jointly is a natural way of handling this. An MCMC sampler where each volatility was treated in a separate step was introduced in Jacquier et al. (1994), and efficient algorithms for multi-state MCMC sampling schemes were suggested by Carter and Kohn (1994) and de Jong and Shephard (1995). Many extensions of the simple stochastic volatility model above have been proposed in the literature, such as correlations between the $\{u_t\}$ and $\{v_t\}$ processes, capturing leverage effects, or fat-tailed distributions for u_t. Inference with these more general models and ways of choosing between them are discussed in Jacquier et al. (2004).

Recently, the focus in finance has shifted more towards continuous-time models, and continuous-time versions of stochastic volatility models have been proposed. In particular, Barndorff-Nielsen and Shephard (2001) introduce a class of models where the volatility behaves according to an Ornstein–Uhlenbeck process, driven by a positive Lévy process without Gaussian component (a pure jump process). These models introduce discontinuities (jumps) into the volatility process. Barndorff-Nielsen and Shephard (2001) also consider superpositions of such processes. Bayesian inference in such models through MCMC methods is complicated by the fact that the model parameters and the latent volatility process are often highly correlated in the posterior, leading to the problem of over-conditioning. Griffin and Steel (2006b) propose MCMC methods based on a series representation of Lévy processes, and avoid over-conditioning by dependent thinning methods. In addition, they extend the model by including a jump component in the returns, leverage effects and separate risk pricing for the various volatility components in the superposition. An application to stock price data shows substantial empirical support for a superposition of processes with different risk premiums and a leverage effect. A different approach to inference in such models is proposed in Roberts et al. (2004), who suggest a re-parameterization to reduce the correlation between the data and the process. The re-parameterized process is then proposed only in accordance with the parameters.

Semi-and Nonparametric Models

The development and use of Bayesian nonparametric methods has been a rapidly growing topic in the statistics literature, some of which is reviewed in Müller and Quintana (2004).

However, the latter review does not include applications to time series, which have been perhaps less prevalent than applications in other areas, such as regression, survival analysis and spatial statistics.

Bayesian nonparametrics is sometimes considered an oxymoron, since Bayesian methods are inherently likelihood-based, and thus require a complete probabilistic specification of the model. However, what is usually called Bayesian nonparametrics corresponds to models with priors defined over infinitely dimensional parameter spaces (functional spaces) and this allows for very flexible procedures, where the data are allowed to influence virtually all features of the model.

Defining priors over collections of distribution functions requires the use of random probability measures. The most popular of these is the so-called Dirichlet process prior introduced by Ferguson (1973). This is defined for a space Θ and a σ-field B of subsets of Θ. The process is parameterized in terms of a probability measure H on (Θ,B) and a positive scalar M. A random probability measure, F, on (Θ,B) follows a Dirichlet process $DP(MH)$ if, for any finite measurable partition, $B_1,\ldots, B_k$, the vector $(F(B_1),\ldots, F(B_k))$ follows a Dirichlet distribution with parameters $(MH(B_1),\ldots, MH(B_k))$. The distribution H centres the process and M can be interpreted as a precision parameter.

The Dirichlet process is (almost surely) discrete and, thus, not always suitable for modelling observables directly. It is, however, often incorporated into semiparametric models using the hierarchical framework

$$y_i \sim g(y_i|u_i) \text{ with } u_i \sim F \text{ and } F \sim \mathrm{DP}(MH), \quad (3)$$

where $g(\cdot)$ is a probability density function. This model is usually referred to as a 'mixture of Dirichlet processes'. The marginal distribution for y_i is a mixture of the distribution characterized by $g(\cdot)$. This basic model can be extended: the density $g(\cdot)$ or the centring distribution H can be (further) parameterized, and inference can be made about these parameters. In addition, inference can be made about the mass parameter M. Inference in these models with the use of MCMC algorithms has become quite feasible, with methods based on MacEachern (1994) and Escobar and West (1995).

However, the model in (3) assumes independent and identically distributed observations and is, thus, not directly of interest for time series modelling. A simple approach followed by Hirano (2002) is to use (3) for modelling the errors of an autoregressive model specification. However, this does not allow for the distribution to change over time. Making the random probability measure F itself depend on lagged values of the variable under consideration y_t (or, generally, any covariates) is not a straightforward extension. Müller et al. (1997) propose a solution by modelling y_t and y_{t-1} jointly, using a mixture of Dirichlet processes. The main problem with this approach is that the resulting model is not really a conditional model for y_t given y_{t-1}, but incorporates a contribution from the marginal model for y_{t-1}. Starting from the stick-breaking representation of a Dirichlet process, Griffin and Steel (2006a) introduce the class of order-based dependent Dirichlet processes, where the weights in the stick-breaking representation induce dependence between distributions that correspond to similar values of the covariates (such as time). This class induces a Dirichlet process at each covariate value, but allows for dependence. Similar weights are associated with similar orderings of the elements in the representation and these orderings are derived from a point process in such a way that distributions that are close in covariate space will tend to be highly correlated. One proposed construction (the arrivals ordering) is particularly suitable for time series and is applied to stock index returns, where the volatility is modelled through an order-based dependent Dirichlet process. Results illustrate the flexibility and the feasibility of this approach. Jensen (2004) uses a Dirichlet process prior on the wavelet representation of the observables to conduct Bayesian inference in a stochastic volatility model with long memory.

Conclusion: Where Are We Heading?

In conclusion, Bayesian analysis of time series models is alive and well. In fact, it is an ever

growing field, and we are now starting to explore the advantages that can be gained from using Bayesian methods on time series data. Bayesian counterparts to the classical analysis of existing models, such as AR(F)IMA models, are by now well-developed and a lot of work has already been done there to make Bayesian inference in these models a fairly routine activity. The main challenge ahead for methodological research in this field is perhaps to further develop really novel models that not merely constitute a change of inferential paradigm but are inspired by the new and exciting modelling possibilities that are available through the combination of Bayesian methods and MCMC computational algorithms. In particular, nonparametric Bayesian time-series modelling falls in that category and I expect that more research in this area will be especially helpful in increasing our understanding of time series data.

See Also

- ARCH Models
- Bayesian Econometrics
- Bayesian Methods in Macroeconometrics
- Bayesian Non-parametrics
- Bayesian Statistics
- Econometrics
- Long Memory Models
- Markov Chain Monte Carlo Methods
- State Space Models
- Statistics and Economics
- Stochastic Volatility Models
- Time Series Analysis

Bibliography

Albert, J., and S. Chib. 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics* 11: 1–15.

Barndorff-Nielsen, O., and N. Shephard. 2001. Non-Gaussian OU based models and some of their uses in financial economics. *Journal of the Royal Statistical Society Series B* 63: 167–241 (with discussion).

Bauwens, L., and M. Lubrano. 1998. Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal* 1: C23–C46.

Bauwens, L., M. Lubrano, and J.F. Richard. 1999. *Bayesian inference in dynamic econometric models*. Oxford: Oxford University Press.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–327.

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

Canova, F. 2004. Testing for convergence clubs in income per capita: A predictive density approach. *International Economic Review* 45: 49–77.

Carter, C., and R. Kohn. 1994. On Gibbs sampling for state space models. *Biometrika* 81: 541–553.

Chib, S., and E. Greenberg. 1994. Bayes inference in regression models with ARMA (p, q) errors. *Journal of Econometrics* 64: 183–206.

de Jong, P., and N. Shephard. 1995. The simulation smoother for time series models. *Biometrika* 82: 339–350.

Durlauf, S., and P. Johnson. 1995. Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics* 10: 365–384.

Engle, R. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1008.

Escobar, M., and M. West. 1995. Bayesian density-estimation and inference using mixtures. *Journal of the American Statistical Association* 90: 577–588.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1: 209–230.

Fernández, C., E. Ley, and M. Steel. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576.

Frühwirth-Schnatter, S., and S. Kaufmann. 2006. Model-based clustering of multiple time series. *Journal of Business and Economic Statistics* 26: 78–89.

Gamerman, D. 1997. *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton: Chapman and Hall/CRC.

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

Geweke, J., and M. Keane. 2006. Smoothly mixing regressions. *Journal of Econometrics* 138: 291–311.

Geweke, J., and N. Terui. 1993. Bayesian threshold autoregressive models for nonlinear time series. *Journal of Time Series Analysis* 14: 441–454.

Granger, C., and R. Joyeux. 1980. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–39.

Griffin, J., and M. Steel. 2006a. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* 101: 179–194.

Griffin, J., and M. Steel. 2006b. Inference with non-Gaussian Ornstein-Uhlenbeck processes for stochastic volatility. *Journal of Econometrics* 134: 605–644.

Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.

Harrison, P., and C. Stevens. 1976. Bayesian forecasting. *Journal of the Royal Statistical Society Series B* 38: 205–247 (with discussion).
Harvey, A. 1981. *Time series models*. Oxford: Philip Allen.
Harvey, A., T. Trimbur, and H. van Dijk. 2006. Trends and cycles in economic time series: A Bayesian approach. *Journal of Econometrics* 140: 618–649.
Hirano, K. 2002. Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* 70: 781–799.
Hsu, N., and F. Breidt. 2003. Bayesian analysis of fractionally integrated ARMA with additive noise. *Journal of Forecasting* 22: 491–514.
Jacobson, T., and S. Karlsson. 2004. Finding good predictors for inflation: A Bayesian model averaging approach. *Journal of Forecasting* 23: 479–496.
Jacquier, E., N. Polson, and P. Rossi. 1994. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics* 12: 371–417 (with discussion).
Jacquier, E., N. Polson, and P. Rossi. 2004. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics* 122: 185–212.
Jensen, M.J. 2004. Semiparametric Bayesian inference of long-memory stochastic volatility models. *Journal of Time Series Analysis* 25: 895–922.
Kleibergen, F., and H. van Dijk. 1993. Non-stationarity in GARCH models: A Bayesian analysis. *Journal of Applied Econometrics* 8: S41–S61.
Koop, G. 2003. *Bayesian econometrics*. Chichester: Wiley.
Koop, G., and S. Potter. 1999. Bayes factors and nonlinearity: Evidence from economic time series. *Journal of Econometrics* 88: 251–281.
Koop, G., J. Osiewalski, and M. Steel. 1995. Bayesian long-run prediction in time series models. *Journal of Econometrics* 69: 61–80.
Koop, G., E. Ley, J. Osiewalski, and M. Steel. 1997. Bayesian analysis of long memory and persistence using ARFIMA models. *Journal of Econometrics* 76: 149–169.
Leamer, E. 1978. *Specification searches: Ad Hoc inference with nonexperimental data*. New York: Wiley.
MacEachern, S. 1994. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics B* 23: 727–741.
Marriott, J., N. Ravishanker, A. Gelfand, and J. Pai. 1996. Bayesian analysis of ARMA processes: Complete sampling-based inference under exact likelihoods. In *Bayesian analysis in statistics and econometrics*, ed. D. Berry, K. Chaloner, and J. Geweke. New York: Wiley.
Müller, P., and F. Quintana. 2004. Nonparametric Bayesian data analysis. *Statistical Science* 19: 95–110.
Müller, P., M. West, and S. MacEachern. 1997. Bayesian models for nonlinear autoregressions. *Journal of Time Series Analysis* 18: 593–614.
Odaki, M. 1993. On the invertibility of fractionally differenced ARIMA processes. *Biometrika* 80: 703–709.
Paap, R., and H. van Dijk. 2003. Bayes estimation of Markov trends in possibly cointegrated series: An application to U.S. consumption and income. *Journal of Business and Economic Statistics* 21: 547–563.
Pai, J., and N. Ravishanker. 1996. Bayesian modeling of ARFIMA processes by Markov chain Monte Carlo methods. *Journal of Forecasting* 16: 63–82.
Phillips, P. 1991. To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* 6: 333–473 (with discussion).
Roberts, G., O. Papaspiliopoulos, and P. Dellaportas. 2004. Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society, Series B* 66: 369–393.
Shephard, N., and M. Pitt. 1997. Likelihood analysis of non-Gaussian measurement time series. *Biometrika* 84: 653–667.
Smith, P.A., and P.M. Summers. 2005. How well do Markov switching models describe actual business cycles? The case of synchronization. *Journal of Applied Econometrics* 20: 253–274.
West, M., and P. Harrison. 1997. *Bayesian forecasting and dynamic models*, 2nd ed. New York: Springer.

Beccaria, Cesare Bonesana, Marchese di (1738–1794)

Peter Groenewegen

Keywords

Beccaria, C. B.; Bentham, J.; Cantillon, R.; Division of labour; Genovesi, A.; Hume, D.; Locke, J.; Montesquieu, C. de; Physiocracy; Population; Productive vs. unproductive labour; Property; Public finance; Quesnay, F.; Smuggling; Taxation; Utilitarianism; Wealth

JEL Classifications

B31

Italian economist, philosopher and statesman, Beccaria was born in Milan in 1738, educated at Parma and in law at Pavia, appointed Professor of Political (Public) Economy or Cameral Science in Milan (1768), resigned his chair to enter public service (1772), where he encouraged and implemented monetary, general economic and

penal reforms and advocated a decimal system of weights, measures and coin. He died in Milan in 1794. Beccaria's greatest fame derives from his *Essay on Crimes and Punishment* (1764a), which made his European reputation almost overnight and ensured his magnificent reception when he visited Paris in 1766. Among others, it exerted considerable influence on Bentham's utilitarian philosophy (Halévy 1928) and popularized the phrase, 'the greatest happiness of the greatest number' (Beccaria 1764a, Introduction). He also enjoyed considerable reputation as an economist. This was based on his work on Milanese monetary problems of 1762 and the outline of his teaching programme and inaugural lecture of 1769 (translated into French and English). His most important economic work is an unfinished treatise, *Elementi di economia pubblica* (written in 1771 but not published till 1804), but his mathematical contribution to the economics of taxation and smuggling (1764b) is also of considerable interest (see Theocharis 1961).

Beccaria (1764b) starts with a methodological point on the use of algebra in political and economic reasoning. He considered such use only legitimate when the analysis concerned quantities, hence not all subject matter of these sciences was amenable to mathematical reasoning. He then illustrates the use of algebra for solving an economic problem, namely, how much of a given quantity of merchandise must merchants smuggle in order to break even, even if the remainder of the goods is confiscated. The essay may have been inspired by Hume's 'Of the Balance of Trade' (1752, p. 76) with its comment on 'Swift's maxim' [that] 'in the arithmetic of the customs, two and two make not four, but often only one', because alterations in rates may alter revenue quite disproportionately.

Beccaria's plan for university instruction in economics and his inaugural lecture develop a classification of the subject matter into five, interconnected parts: general principles and overview, agriculture, trade, manufactures and public finance. Further subdivisions into chapters are reminiscent of the table of contents of Cantillon (1755), a work he appears to have studied closely, though the historical part of his inaugural lecture only acknowledges Vauban, Melon, Montesquieu, Uztariz, Ulloa, Hume and Genovesi. The last is described as the father of Italian economics (Beccaria 1769). Groenewegen (1983) demonstrates that Beccaria's economic sources also included Locke and Quesnay's articles published in the French *Encyclopédie*. The last gave parts of the *Elementi* a Physiocratic flavour; for example, in the analysis of large- and small-scale farming, productive and unproductive labour and, more generally, its emphasis on the importance of agriculture.

Beccaria sees political economy as a highly practical subject, because it is part of the science of legislation and politics. Its purpose is to 'increase the wealth of the state and its subjects, by giving instruction on the most appropriate and useful management of the national revenue and that of the sovereign' (1769, p. 341). Although abstract treatment of the science is therefore largely rejected as inappropriate for such a practical subject, Beccaria maintains that serious discussion of its elements needs an introduction of general principles. A definition of wealth as 'things not only necessary but also convenient and elegant', starts these principles in Part *I* of the *Elementi*. Because wealth consists of goods designed to meet the needs of food, shelter and clothing, the science can be justifiably subdivided into parts derived from the sectors of production and exchange which supply the various wants of mankind. Raw materials are drawn from farming, pastoral activity, mineral exploitation and fishing, hence agriculture is the first part of political economy. Raw materials require work and preparation before they can be used, hence manufacturing is the second part. Efficient production of wealth creates a surplus available for exchange, hence commerce including value, money and credit constitutes the third part to be treated. Since protection of property is a prerequisite for efficient production and trade, public finance explaining how these expenses of government are met is the fourth element. Finally, Beccaria suggests a fifth topic to cover police and other government activity, but nothing of this nor the public finance part of his *Elementi* were ever completed. Having defined the scope of the subject in terms of wealth and the component parts helping its production,

Beccaria elaborates on the principles in his theory of reproduction, or the combination of labour, time and capital which ensures the continuation of production activity. Here Beccaria demonstrates awareness of the links between division of labour and trade and recognizes that the prices which circulate commodities are regulated by necessary costs of production. A general analysis of the cost of labour or wages, of the advances and other means of production and of those incurred by the state in its essential protection of production activity, is therefore required. Beccaria further develops these general principles by examining the nature and interdependence of work and consumption, introducing considerations of thrift, value, profit, useful work, variability of wants and difficulties in measuring the subsistence wage of workers. A discussion of the principle of population concludes the analysis of the 'simple truths' and 'self-evident axioms' from which the whole science of political economy can be deduced, as Beccaria intended to demonstrate in the other parts of his work. Of these, the completed chapters in Part IV on value, money and exchange are of the greatest interest.

Selected Works

1764a. An essay on crime and punishment. *London: J. Almon, 1767.*

1764b. An attempt at an analysis of smuggling. *II Caffè,* vol 1. Brescia, 118–19. In *Precursors in mathematical economics*: *An anthology*, ed. W. Baumol and S. Goldfield. London: London School of Economics, 1968, 149–50.

1769. *A discourse on public oeconomy and commerce*. Translated from the Italian. London: J. Dodsley.

1771. Elementi di economia pubblica. In *Cesare Beccaria Opere,* ed. S. Romagnoli. Florence: Sansoni. 1958.

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. Reprint with English translation edited by H. Higgs for the Royal Economic Society. London: Macmillan. 1931.

Groenewegen, P.D. 1983. Turgot, Beccaria and Smith. In *Italian economics past and present*, ed. P.-D. Groenewegen and J. Halevi. Sydney: Frederick May Foundation for Italian Studies.

Halévy, E. 1928. *The Growth of Philosophic Radicalism*. Trans. M. Morris. Boston: Beacon Press. 1955.

Hume, D. 1752. Of the balance of trade. Reprinted in *David Hume, writings on economics*, E. Rotwein. London: Nelson. 1955.

Theocharis, R.D. 1961. *Early developments in mathematical economics*. London: Macmillan.

Becker, Gary S. (Born 1930)

Casey B. Mulligan

Abstract

Gary S. Becker has produced major economics books and articles for more than 50 years. His studies dominate labour economics and have significantly impacted studies of crime, habit formation, and other important behaviours once considered beyond the scope of economics. Some of Gary's lasting impact can be attributed to abstraction from institutional detail and his 'thinking problems through fully'.

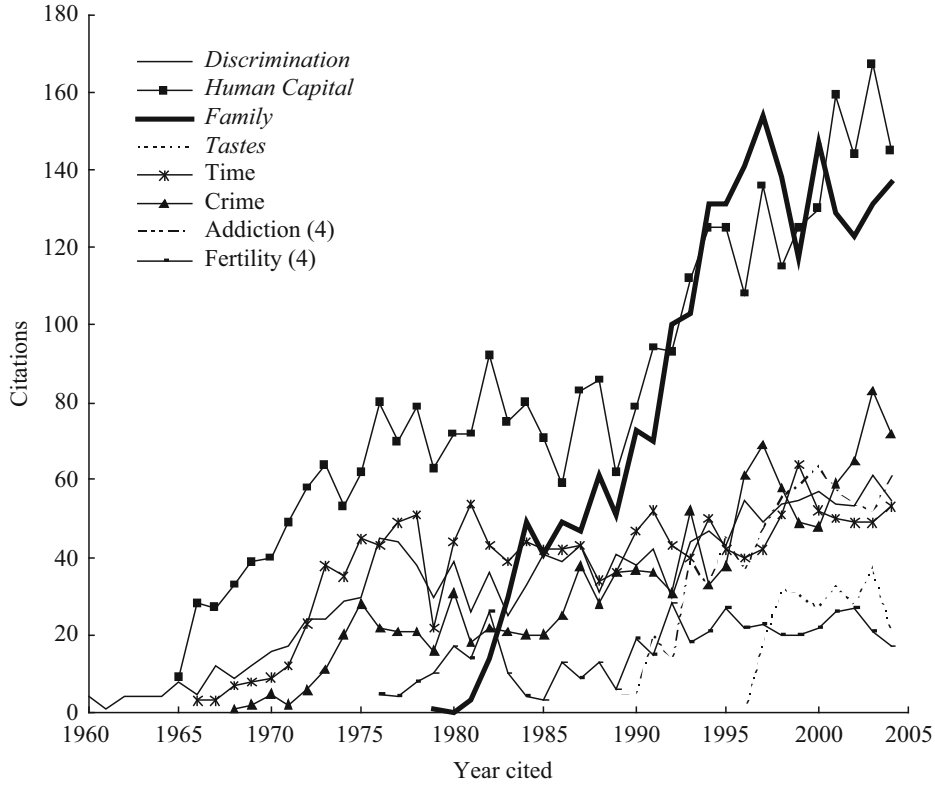
Keywords

Becker, G; Buchanan, J; Chicago School; Downs, A; Family economics; Flat tax; Friedman, M; Human capital; Knight, F; Labour market discrimination; Politics and economics; Pressure groups; Ricardian equivalence theorem; Rotten kid theorem; Schumpeter, J; Stigler, G; Time allocation; Tullock, G

JEL Classifications

B31

I walk over to my collection of *The American Economic Review,* and pick up the very first (and now disintegrating) issue, dated 1952, and notice an article entitled 'A Note on Multi-Country Trade'. Its author is Gary S. Becker. By the time you read this, you probably can pick up the very latest issue



Becker, Gary S. (Born 1930), Fig. 1 Citations of Becker's major books and articles, excluding those on political economy
Note: For Addiction and Fertility, I sum citations for the four articles in each class; some double counting may occur due to articles that cite more than one of the four. Becker's political economy articles may be more important than the Fertility and Addiction articles, but for clarity the former are omitted from Fig. 1 and deferred until later. Social Economics (Becker and Murphy 2003b) is also omitted because, as of 2005, its annual citations were fewer than ten

from your collection and find an article by Gary S. Becker! If you did so in 2005, I can guarantee it: the article was entitled 'The Quality and Quantity of Life and the Evolution of World Inequality'. Gary published an important article in the very first issue of the *Journal of Law and Economics*, 'Competition and Democracy' (1958). He published an article, 'Deadweight Costs and the Size of Government' in the 46th volume of the same journal (Becker and Mulligan 2003a); it may have the same potential, although I must admit that its importance cannot yet be judged impartially.

Figure 1 quantitatively examines Gary's work over a half century. The vertical axis measures, from the Social Science Citation Index (SSCI), the number of articles citing each of Gary's books and major research projects. Each citation has a citer and a citee. The citees are Gary's *Economics of Discrimination* (1957, various editions), *Human Capital* (1964, various editions), *A Treatise on the Family* (1981, various editions), *Accounting for Tastes* (1996), 'A Theory of the Allocation of Time' (1965), 'Crime and Punishment: An Economic Approach' (1968), four journal articles on addiction (Becker and Murphy 1988a; Becker et al. 1991, 1994; Becker 1992), and four journal articles on fertility (Becker and Lewis 1973; Becker and Tomes 1976; Becker and Barro 1988; Barro and Becker 1989). The citers are social science journal articles published in the year indicated on the horizontal axis. Since the articles are typically peer-reviewed and the journals are academic, the vertical axis is a measure (admittedly imperfect) of how important

Gary's various works were in making intellectual progress, or in shaping the thinking behind intellectual progress, in social science. Notice the scale on the vertical axis – it reaches past 100 citations per year per work of Gary's – and remember that there are tenured professors at leading economics departments whose citations *combined for all of their works and all of their lives* do not reach these levels. Also notice the scale on the horizontal axis: it begins in 1960. (A fuller analysis of citations would separate year effects from other determinants of citations – for example, the number of journals covered by SSCI may increase over time; I owe this point to Bill Landes. However, the reader might make some guess at the year effects from the fact that *Human Capital's* citation time series is quite similar to those of Schumpeter 1942, and Downs 1957. *Human Capital's* citations significantly exceed and grow faster than those of Friedman 1957, and Friedman and Schwartz 1963.) *Discrimination* and *Treatise* are both heavily cited, but their first editions appeared 24 years apart. The addiction work first appeared 31 years after *Discrimination*. (The two pressure group papers, discussed later, appeared 26 and 28 years respectively after *Discrimination,* and surpassed 50 combined citations per year by 1990.) If Gary manages another big hit during the next few years, that would be a 50-year span.

In 1999 – to me that seems a long time ago – I visited Wayne State University and met for the first time John Owen, a labour economist whom I knew by reputation. I was both flattered and wiser for this emeritus professor's making the trip to campus to meet me and hear my seminar. As we talked, his style of economic reasoning seemed familiar to me, so I asked him where he obtained his Ph.D. He replied 'I am one of Gary's students, of course'. Apparently Gary Becker alumni have been filling the emeritus professor ranks for a while now. Jack Nicklaus had better win the Masters a couple more times if he wants to be as good at golf as Gary is at economics.

It could be a hundred years or more before economics sees another iron man like Gary. Biographies about Becker should be written if for no other reason than people will ask 'How did he do it?' But why should I be writing a biography, and what could I possibly contribute to answering this difficult question? After all, Gary is closer in age to my grandfather than to my father, so I am certainly no authority on where he was born, what kind of student he was, and so on. By the time I first met Becker in 1991, his Nobel Prize was only one year away. On the other hand, I do know (some more closely than others) many of the important intellectual companions in his life, including Guity Nashat Becker, Aaron Director, Milton Friedman, Jacob Mincer, Sherwin Rosen, Gale Johnson, Jim Coleman, Bill Landes, Bob Lucas, Sam Peltzman, Dick Posner, Isaac Erlich, Kevin Murphy, Robert Barro, Eddie Lazear, Victor Fuchs, Ed Glaeser, Andy Rosenfield, and Tomas Philipson. The opportunity cost of time is certainly lower for me than for those on this list. (Becker's work is so widely applicable that it can even be used to predict who'd write his biography(ies).) Gary loves economics dearly, so perhaps my best tribute would exploit my perspective as a 14-year student, colleague, and friend of Gary's – who was always glad to hear stories about Gary's achievements and the University of Chicago from older students and colleagues such as John Owen and the other names mentioned above – in order to convey some information about Gary's life that is not readily found in a literal reading of his published work, and might help future economists progress a little faster.

The first section raises the question of whether and how the University of Chicago might have affected Gary's intellectual contributions. The second section discusses Gary's timing in the marketplace for economics ideas. Did Gary leave some potential unrealized? The third section addresses this question, with emphasis on economic approaches to political behaviour. Gary's results sometimes seem pretty obvious, but the fourth section explains how this judgement is usually the perspective of hindsight. It offers a number of remarkable examples of how economists, including Gary himself, took a while to fully understand the implications of his economic approach to the family, the labour market, and other areas.

Did Chicago Matter?

I'm told that Becker first came to the University of Chicago in 1951 as a graduate student. How much did it matter that he came to Chicago rather than accepting a nice fellowship at Harvard? Some of Gary's undergraduate work at Princeton foreshadowed two of his important contributions to economics. First was the trade paper I mentioned above. Trade theory features prominently in *The Economics of Discrimination,* and even today is still an intense interest of Gary's, as his colleagues today can see any time a trade paper is presented in front of the economics faculty. I doubt that Chicago has done much to cultivate this interest. Second is Gary's 'A Theory of Competition among Pressure Groups for Political Influence' (1983). In one sense, Chicago was necessary for the production of this paper, because it grew out of a comment on Peltzman's 1976 paper in the *Journal of Law and Economics* and a dialogue with Stigler as to whether the political process favoured efficiency or special interests. However, Gary may have been thinking seriously about competition in the public sector during his Princeton days, since already in his first year at Chicago he was writing the first drafts of 'Competition and Democracy', which was published in the inaugural issue of the *Journal of Law and Economics* only after being squashed at the *Journal of Political Economy* by another important Chicago economist, Frank Knight. (Today Gary credits some of his early thinking on democracy to his reading of Schumpeter's *Capitalism, Socialism, and Democracy,* 1942, but he does not remember whether he read it before coming to Chicago, or shortly after.)

Before coming to Chicago, Gary was already dissatisfied with the lack of applications of economics to important social problems, although his Princeton work does not yet show any success at resolving his discontent. Perhaps Chicago, and especially Milton Friedman, inspired or at least encouraged the application of economics beyond the usual areas. As Gary says, '[Friedman] emphasized that economic theory was not a game played by clever academicians, but was a powerful tool to analyse the real world. His course was filled with insights both into the structure of economic theory and its application to practical and significant questions' (Becker 1993). Gary is now known for his application of economic theory to practical and significant questions, from time allocation and fertility to inequality and addictions.

Gary sometimes explains, 'I was such an outsider from the eastern and western establishments for so long'. Universities like Stanford, Harvard, and Yale have never showed any interest in hiring him, although Harvard granted him an honorary degree in 2003. Gary's abilities as an economist are so extraordinary that, despite being an outsider, and having such a large fraction of his productivity ahead of him, he was recognized in 1967 by the American Economic Association as the best young economist at the time (he won the their John Bates Clark Medal in that year). Gary's outsider position would have been different had he turned down Chicago's fellowship, but fortunately for him citations and academic job offers have very different production functions, at least as regards their use of personal acquaintances as inputs.

At Chicago Gary met, loved, and improved the workshop system. Columbia University was the first beneficiary of those improvements when he and Mincer created the Labor Economics workshop (Landes 1998). Gary started a workshop when he returned to Chicago in 1970, which for many years was co-organized with Sherwin Rosen, and is now affectionately known as the 'Applications Workshop'. By the time I began attending economics workshops in 1991, practically all had become (and maybe had always been?) something like lecture series, and were a form of *output* of the idea production process, namely, a process for disseminating finished research results. But the Applications Workshop was and is deliberately different; research papers are invited in their infancy, and 85 of the 90 minutes consist of the audience's (especially Gary's; Gary carefully reads the paper beforehand) trying to push the author in new and better directions. Students regularly come to the workshop to hear what Gary has to say, and, in the midst of a graduate programme that can easily

overwhelm them with technical detail, learn that good choices of research question and basic strategy for seeking an answer are important and scarce academic skills. Gary later organized with the late James Coleman (Richard Posner continues the tradition) an interdisciplinary workshop on applications of rational models to economics, sociology, law, politics, anthropology, and so on. The success of these two workshops make Chicago a unique and highly stimulating experience for faculty and students, and probably would not be possible if it weren't for Gary's extraordinary breadth of knowledge, quickness of mind, and insatiable appetite for workshops.

Abstracting from Institutional Detail

The workshop system and Economics 301 (Chicago's first Ph.D. course in price theory) were important means by which Gary received his inheritance from Chicago, and made his bequest to students at Chicago and Columbia, where Becker was an economics professor from 1957 to 1970. I mentioned Friedman's lesson that economic theory was not a game played by clever academicians, but was a powerful tool to analyse practical and significant questions. Chicago was methodologically unique in two other ways. Despite their working on practical questions, Chicago economists were willing, and even eager, to abstract from institutional details, and view price theory as a general method to understanding many different behaviours. This approach was particularly novel in labour economics, where labour unions, marriage bars, and other personnel practices were often interpreted as having an independent influence on labour market outcomes, rather than as outcomes themselves of more basic and ubiquitous forces. Columbia's Jakob Mincer also practised this methodology in his enduring work on labour supply (for example, Mincer 1962). Labour market institutions like trade unions and monster.com (an internet site where employers can read resumes posted by potential employees) come and go, but the fundamental economic forces include the income and substitution effects on labour supply featured at Columbia by Mincer and at Chicago by Lewis (1956), and are an important part of explanations of why labour market outcomes vary over time and across regions. It's no coincidence that Becker and Mincer together created the Labor Economics workshop at Columbia, and work appearing during these years by Becker, Mincer, and students continued the practice. (William Landes – Gary's student, colleague and friend during both the Chicago and Columbia years – wrote in 1998 an excellent biography of Becker which explains more about the Columbia days and Gary's influence on the law and economics field. To Landes' account I would add that Gary still credits the City of New York with inspiring 'Crime and punishment'. One day he illegally parked his car near Columbia's campus because he calculated it to be more important to attend a dissertation defense than to avoid the city's illegal parking fine.)

Human Capital also has some roots in Gary's time at Chicago before 1957. Chicago's agricultural economics group (Gary was one of the participants in those days), especially Ted Schultz, had attributed much of the underdevelopment problem to a lack of human capital investment. Gary's *Human Capital* explains why some people have more income from employment than others by viewing labour income as a dividend on historical investments, which in turn are understood as particular instances of capital accumulation. The basic concepts do not include labour market institutions, but rather the time value of money, ageing, the allocation of time, and other determinants of the costs and benefits of enhancing a person's productivity in the marketplace. Becker's abstractions facilitated applications of human capital theory beyond (perhaps) even what he had anticipated, including the determinants of sickness and health (Grossman 1972, a Columbia Ph.D. student 1964–70), and the evolution of species (Robson and Kaplan 2003).

'A Theory of the Allocation of Time' introduced the concepts of 'full income' and the 'full price' of a commodity. A commodity's full price combines the expenditures of money and time required to acquire one unit of its services. Because households differ in terms of the opportunity cost of their time, and perhaps also their time-efficiency in obtaining commodities, they will face different full prices even though they

face the same money price. For example, the substitution effect suggests that richer households (to the extent that the market rewards them highly for their time) would have fewer children and, per unit consumed, would replenish less often their inventories of household commodities (and currency: Karni 1973). (For the same reason, Gary is perennially puzzled why rich people play golf; he plays tennis.) Full income is the money income that would be obtained if time were allocated in order to maximize money income. In many ways, full income permits time allocation to be studied as a particular application of consumer demand, because full income is spent on some combination of market expenditures on commodities and implicit expenditures on non-work time. Full income and full price are not institution-specific concepts, permitting 'Time' to be applied in so many different sub-fields, including monetary economics, fertility, lobbying (Mulligan and Sala-i-Martin 1999), altruism (Mulligan 1997), and even Communism (Boycko 1992).

Public Policy Schisms

Milton Friedman's *Capitalism and Freedom* (1962) and *Free to Choose* (1981) clearly advertise the view that inefficient public policies are bad ideas unfortunately and inexplicably hatched by policy-makers, which can be rectified merely by giving some combination of voters, politicians, and bureaucrats a better economic education. If Gary continued that tradition, as with his *Business Week* column and internet blog, he did so with much less vigour. One of Chicago's important influences on Gary came from George Stigler, who often viewed public policies as the rational choices of politicians and the people who can influence them. Perhaps Stigler's influence was stronger because Friedman was there to contrast it, but in any case it's hard to see any Friedman in 'Pressure Groups' (1983) or 'The Family and the State' (1988b).

Interestingly, this schism persists today in Chicago's Economics Department and the economics profession more widely. A public finance group, embodied at Chicago in its macro group (for example, Lucas and Stokey 1983; Shimer and Werning 2003), aims at technical and normative public policy improvements, whereas political economists (for example, Becker and Murphy 1988; Mulligan et al. 2004) view public policies and their imperfections as the outcomes of other economic forces, such as demography, political competitiveness, and the technology of tax collection.

Becker (1983) also tries to bridge a gap among political economists – a gap defined according to whether they see special interests or efficiency as the primary determinant of actual public policies. He points out that a huge number of groups would like special favours from the government, but only a few can ultimately be successful. These groups compete with each other to obtain the favours. All else the same, groups advocating efficient public policies have an advantage because (by definition of efficiency) their policy proposals would hurt relatively little. Of course, group cohesion, political entry barriers, group size, and other variables may give particular groups an intrinsic advantage, but the competitive activity of special interest groups helps deliver efficient policies to the public sector rather than crowding out such policies with inefficient special favours.

Unfortunately, Becker has not (yet) bridged another gap among political economists – a gap defined by the degree of attention to institutional detail. It's interesting that labour economics work done by Gary and others at Chicago is praised for its lack of institutional detail (detail now considered unnecessary for understanding the major economic forces at work), whereas the political economics work is criticized, at least so far, for the same lack of detail.

Timing in the Marketplace for Ideas: Human Capital or Luck?

Human Capital and 'Time' had some good fortune in their timing, both in terms of the ultimate demand for these ideas and in terms of the supply of intellectual building blocks. For example, *Human Capital*'s citations accelerated in the late 1980s as the profession came to realize the

important wage structure changes that were occurring and began to write about them; human capital theory is probably the most common way of organizing and interpreting such observations. It may also be fortunate that, since 1940, the Census Bureau has been asking more people more questions about wages and schooling than about household expenditure, hence stimulating more empirical research on wages and schooling than empirical research on consumption.

Perhaps there was also good fortune on the input side. Mincer was making significant progress in the empirical analysis of labour supply and the empirical analysis of wage determinants. The economics of consumption was a very lively subject at Chicago in the 1950s, as evidenced by Friedman's *A Theory of the Consumption Function* (1957), work by Margaret Reid (1957), and the beginnings of Chicago's workshop system by Chicago's agricultural economics group. Gary's work on the value of time and life-cycle profiles must have been stimulated in this environment, in part because labour supply and human capital accumulation are such natural applications of the life-cycle way of thinking already apparent in *A Theory of the Consumption Function.* Remember also that Friedman (1957) was preceded by *Income From Independent Professional Practice* (1945), which straddled the fields we would now call consumption and labour economics. (Gregg Lewis was probably yet another Chicago influence in these days.) The economic concept of 'full income' first appeared in 'Time', where Becker credits the phrase to a conversation with Milton Friedman.

Gary adopted and improved the analytical style of *A Theory of the Consumption* Function and the methodology of positive economics more generally. Some consider Friedman's *A Theory of The Consumption Function* the best economics book since the 19th century, and perhaps earlier, because of its convincing and systematic applications of economic theory to important questions. But *Human Capital* may be even better. Both books clearly aim to develop refutable empirical implications from their theories, but *Human Capital* probably does more to help its reader distinguish the important implications from the secondary ones. Gary always advises his students and colleagues to 'think a problem through *fully'* and apparently he followed his own advice in *Human Capital.* Not only is the importance of the basic 'human capital' concept appreciated several decades later, but modern analysis of the labour market still displays more detailed similarities, including attention to specific versus general human capital, comparisons between financial and human capital rates of return, the distinction between the forgone earnings and tuition components of human capital acquisition costs, and so on. Friedman's basic concept of permanent income and the details of his analysis of it (such as 'distributed lags') are less prevalent today, having been displaced by consumption Euler equations. (Almost immediately after *Human Capital*'s publication, its citation flow exceeded and grew faster than that of *A Theory of the Consumption Function.)* By thinking through the problem fully, Gary had produced in the early 1960s an analysis that would depreciate slowly, and thereby still be available in the 1980s to take advantage of the real-world events that drew attention to human capital questions.

Unrealized Potential?

Only people who know Gary personally would know, or dare to believe, that he may have some regrets that he did not realize his full potential. His political economics work is an important instance. He regrets the obscurity of 'Competition and democracy', which has been cited only 33 times – less than once per year. He partly blames editor Aaron Director for forgetting to request revisions or proofs of the manuscript, and himself for not following up on work that he knew to be incomplete.

Political economics research has proliferated since the mid 1980s. Gary feels that progress might have been more significant if 'A Theory of Competition Among Pressure Groups for Political Influence' had received more attention. I am inclined to agree (Mulligan et al. 2004), but it

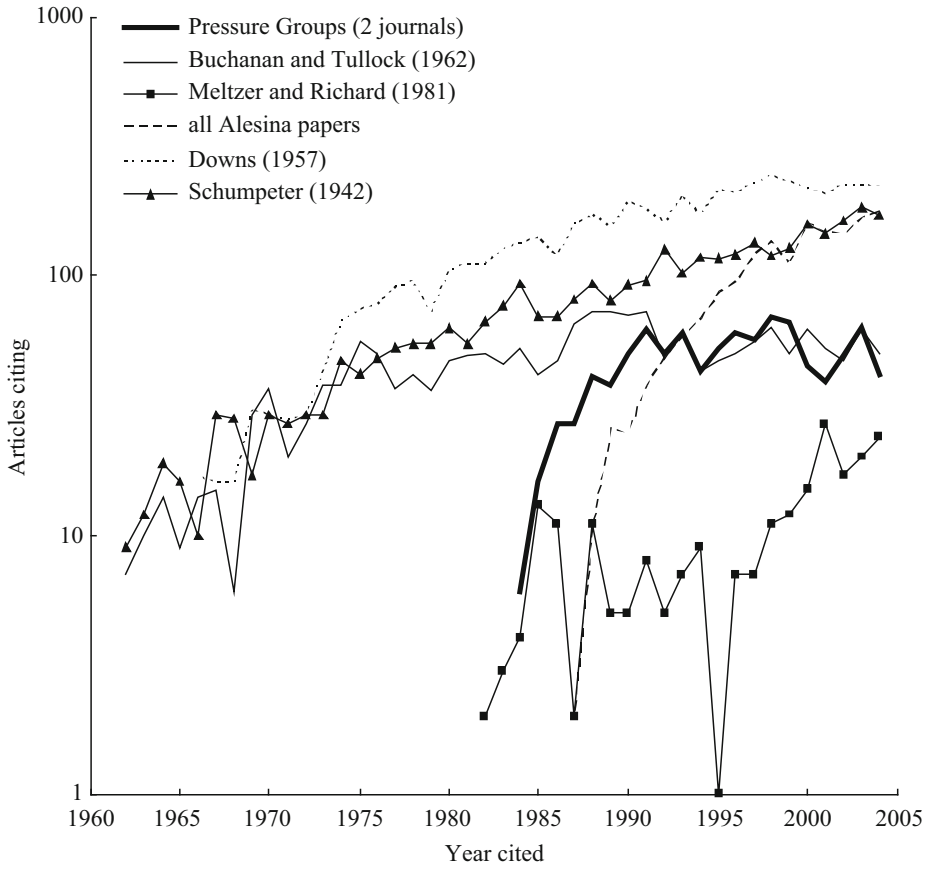
would be much too extreme to say the article was 'ignored'. Yes, it was rejected by the *American Economic Review* and perhaps another journal (Gary does not remember). Nevertheless, it may ultimately be the most cited article appearing in the *Quarterly Journal of Economics* since 1983. It has been cited almost 50 times every year since the 1980s. Only three articles – which happen to be from the economic growth literature: Summers and Heston (1991), Barro (1991), and Mankiw et al. (1992) – have been cited more than 50 times per year for more than a couple of years, and their citation flows have regressed back to Gary's since 2000. (I thank Andrei Shleifer for suggesting comparisons between Becker 1983, and other top *QJE* articles.) Two other *QJE* articles – Katz and Murphy (1992) since 1997 and Fehr and Schmidt (1999) since 2003 – enjoy about the same citation flow as Gary's, but over a much more recent period of time.

'Pressure Groups' citations are in the stratosphere in the universe of journal articles, but nevertheless it has been losing political economics market share as its annual cites have been pretty steady at 50 while the political economics literature has exploded. Figure 2 compares 'Pressure Groups' citations (summed here for the *QJE* and *Journal of Public Economics* articles) with some other political economics work. This time citations are displayed on a log scale. 'Pressure Groups' citations are shown as a thick solid line. Buchanan and Tullock's *Calculus of Consent* (1962) – maybe Nobel Laureate Buchanan's best known work – has had the same flow of citations since 1990, although of course the *Calculus of Consent* was published much earlier and deserves enormous credit for introducing to economics the principle of modelling policy-makers as self-interested. Perhaps more striking is the fact that 'Pressure Groups' citations have not grown with the political economics literature since 1985. For example, Alberto Alesina now accumulates about 200 citations per year (of all of his papers combined, see the dashed line in Fig. 2). Meltzer and Richard's (1981) paper is actually older than Gary's, but it received very few citations until the late 1990s, when its citation flow increased by almost an order of magnitude. Downs (1957, dotted line) and Schumpeter (1942, circles) have also benefited from the growth of political economics. (Mancur Olson's *Logic of Collective Action,* 1965, might have been included in Fig. 2; since 1980 its citation flow is about 50 per cent more than Downs's.)

Perhaps 'Pressure Groups' should have been part of, or led to, a Becker political economics book that worked more fully through the implications of competition for the supply of public policies. Does it matter whether competition is time-intensive or goods-intensive? How competitive are authoritarian regimes? To judge from Gary's treatment of labour economics questions, it seems very likely that a Becker political economics book would have treated fundamental economics forces like deadweight costs, competition, and the allocation of time with little attention to institutional detail. Would such a book have succeeded in the current marketplace for political economics ideas? On the one hand, the answer seems to be 'no' because the current literature prides itself on its analysis of those details; Persson and Tabellini (2004, p. 76) explain, '. . .the devil is in the details, especially the details of electoral systems'; see also Besley and Case (2003, p. 11). On the other hand, Gary's book may have pushed, or at least nudged, the literature in a different direction.

Wasn't It All Obvious?

Perhaps this is a slight exaggeration, but some of Gary's results have been criticized as being too obvious, or adding too little value to simpler non-economic models or common-sense interpretations. I have to admit that I sometimes found it easier to remember the basic results of Gary's journal articles, and to produce simple derivations of my own (for example, Mulligan 1997, ch. 3), than to follow Gary's published derivations. (I don't remember the derivations presented in Gary's University of Chicago courses to be so clear, either. But maybe I deserve much of the blame here; I am much better at following a geometric proof than an algebraic one, whereas Gary



Becker, Gary S. (Born 1930), Fig. 2 Citations of major political economy works

seems to prefer the latter.) To some extent, these critiques have the advantage of hindsight; it is quite normal for original ideas to be expressed later by followers in simpler terms, after a period of what Gary calls 'cleaning up'. However, I believe that Gary's books are easier to follow than several of his journal articles, because the process of writing a whole book was complementary with some cleaning up on his own. This is also part of the reason why Becker and Murphy make such a good team; one of Murphy's extraordinary talents is to quickly conceive of a concise mathematical expression of a new economic idea.

Becker and Tomes (1979, 1986) reinterpret inter-temporal consumption theory and combine it with human capital theory to form a theory of the evolution of inequality from one generation to the next. In the model, altruistic parents allocate dynastic resources between themselves and their children. The opportunities for doing so depend on the process of monetary inheritance (for example, inheritance taxes) and on the technology for investing in the human capital of children. The model predicts that earnings regress to the mean across generations because ability, talent, and so forth (which determine the rate of return to human investment) regress to the mean. Perhaps the most explicit form of the 'too obvious' criticism appeared as Goldberger's (1989) contention that this approach to inheritance is an excessively complicated way of saying 'economic characteristics regress to the mean'. Becker's (1989) reply lists some implications that are more than regression to the mean, although in some cases I think the results still derive from statistical rather then economic modelling assumptions (see Mulligan

1997, and the references cited therein). Nevertheless, Becker's 'micro-economic-optimizing approach' is the only one, to my knowledge, predicting that consumption would regress to the mean more slowly than earnings. It's a nice bonus that, so far, the empirical evidence seems to support Gary in this regard.

For many years, and perhaps even now, it was far from obvious that wages are largely determined by human capital, as evidenced, for example, by the various debates on wage gaps by industry, race, and gender. The opponents of the human capital interpretation of industry gaps have, after several years, softened their view. Gender and race gaps are sometimes attributed to discrimination (Gary gets some credit under this interpretation, too), although there seem to be steady streams of new evidence showing that the effects of human capital have been too quickly misinterpreted as effects of discrimination (see, for example, Smith and Welch 1989; Neal and Johnson 1996, on race gaps, and Mulligan and Rubinstein 2005, on gender gaps).

As Gary began working on the family, he found 'redistribution of income among members does not affect the consumption or welfare of any member because it simply induces offsetting changes in transfers from the head. As a result, each member is at least partially insured against disasters that may strike him' (Becker 1974, p. 1091). Put this way, the result seems obvious. However, the result could not have been fully understood at the time – otherwise the rotten kid theorem, the Ricardian equivalence result, and a number of other results would not have shaken the profession so much. Indeed, Gary himself did not fully appreciate its implications, because he admits not foreseeing how the macroeconomics of fiscal policy would change after 1974 thanks to Barro's (1974) article in the same issue of the *Journal of Political Economy.* (Barro's focus at the time was probably contemporaneous work on fiscal policy, such as Feldstein's famous 1974 article in the previous *JPE.* Barro 1998, explains how the links between Ricardian Equivalence and the Rotten Kid Theorem began to be appreciated only when the *JPE* began preparing the November 1974 issue in which the two articles were to appear.) Peter Diamond's reaction (as reported second-hand by Barro 1998) demonstrates the fallacy of dismissing these results as obvious, '[Ricardian equivalence is] obvious, of no practical significance, and surely not worth . . . research time.' Professor Diamond was giving this advice in 1967 to student Bob Hall, who, if it weren't for his listening, was on the verge of scooping both Becker and Barro.

During the 1996 US presidential campaign, Republican primary candidate Steve Forbes revitalized the idea of replacing the current income tax with a 'flat tax': a tax with no deductions and low marginal rates. I was concerned that a painless tax would be a tax that Congress would exploit to obtain ever larger amounts of revenue, but to me this point was just something clever to publish in the op.-ed. pages or to make people pause at cocktail parties. I vividly remember mentioning this to Gary in March 1996. He was a flat tax fan at the time (see Becker et al. 1996), and told me 'I'm not sure how you would analyse that formally and, besides, Hong Kong refutes your hypothesis: they have a flat tax and a small government'. A few days later he apparently saw the empirical evidence differently, and was excited enough to interrupt his trip in France to type a short first draft of our 'Deadweight Costs and the Size of Government' and attach it to an e-mail to me back at the University of Chicago. By then he was sure how to analyse it: using a simple version of his 1983 pressure group model. The lesson for the young assistant professor: think a problem through *fully*, regardless of how obvious the answer might seem at first glance. The rewards in this case were, among other things, a consistent analysis of tax reforms, spending reforms, and 'flypaper effects' (the tendency of governments to spend non-tax revenue rather than refund it to taxpayers), and a better understanding of the relations between democratic and authoritarian public sectors.

See Also

▶ Family Economics
▶ Human Capital

Selected Works

1952. A note on multi-country trade. *American Economic Review* 42: 558–568.

1957. *The economics of discrimination.* Chicago: University of Chicago Press.

1958. Competition and democracy. *Journal of Law and Economics* 1 105–109.

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

1965. A theory of the allocation of time. *Economic Journal* 75: 493–508.

1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

1973. (With H. Lewis.) On the interaction between the quantity and quality of children. *Journal of Political Economy* 81 (Part II): S279–S288.

1974. A theory of social interactions. *Journal of Political Economy* 82 1063–1093.

1976. (With N. Tomes.) Child endowments and the quantity and quality of children. *Journal of Political Economy* 84 (Part II): S143–S162.

1979. (With N. Tomes.) An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy* 87: 1153–1189.

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

1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 98: 371–400.

1985. Public policies, pressure groups, and deadweight costs. *Journal of Public Economics* 28: 329–347.

1986. (With N. Tomes.) Human capital and the rise and fall of families. *Journal of Labor Economics* 4: S1–S39.

1988. (With R. Barro.) A reformulation of the economic theory of fertility. *Quarterly Journal of Economics* 103: 1–25.

1988a. (With K. Murphy.) A theory of rational addiction. *Journal of Political Economy* 96: 675–700.

1988b. (With K. Murphy.) The family and the state. *Journal of Law and Economics* 31: 1–18.

1989. On the economics of the family: Reply to a skeptic. *American Economic Review* 79: 514–518.

1989. (With R. Barro.) Fertility choice in a model of economic growth. *Econometrica* 57: 481–501.

1991. (With M. Grossman and K. Murphy.) Rational addiction and the effect of price on consumption. *American Economic Review* 81: 237–241.

1992. Habits, addictions, and traditions. *Kyklos* 45: 327–345.

1993. Gary S. Becker – Autobiography. In *Les Prix nobel. The nobel prizes 1992,* ed. T. Frängsmyr. Stockholm: Almqvist and Wiksell International. Online. Available at: http://nobelprize.org/economics/laureates/1992/becker-autobio.html. Accessed 2 Dec 2005.

1994. (With M. Grossman and K. Murphy.) An empirical analysis of cigarette addiction. *American Economic Review* 84: 396–418.

1996. *Accounting for tastes.* Cambridge, MA: Harvard University Press.

1996, 22 February. (With others.) The flat-tax: 'nutty' it's not. *Wall Street Journal.*

2003a. (With C. Mulligan.) Deadweight costs and the size of government. *Journal of Law and Economics* 46: 293–340.

2003b. (With K. Murphy.) *Social economics: market behavior in a social environment*. Cambridge, MA: Belknap Press.

2005. (With T. Philipson and R. Soares.) The quality and quantity of life and the evolution of world inequality. *American Economic Review* 95: 277–291.

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Barro, R. 1991. Economic growth in a cross section of countries. *Quarterly Journal of Economics* 106: 407–443.

Barro, R. 1998. Reflections on Ricardian equivalence. In *Debt and deficits: An historical perspective*, ed. J. Maloney. Cheltenham: Edward Elgar.

Besley, T., and A. Case. 2003. Political institutions and policy choices: Evidence from the United States. *Journal of Economic Literature* 41: 7–73.

Boycko, M. 1992. When higher incomes reduce welfare: Queues, labor supply, and macro equilibrium in

socialist economies. *Quarterly Journal of Economics* 107: 907–920.
Buchanan, J., and G. Tullock. 1962. *The calculus of consent*. Ann Arbor: University of Michigan Press.
Downs, A. 1957. *An economic theory of democracy*. New York: Harper.
Fehr, E., and K. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114: 817–868.
Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82: 905–926.
Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.
Friedman, M. 1962. *Capitalism and freedom*. Chicago: University of Chicago Press.
Friedman, M., and R. Friedman. 1981. *Free to choose*. New York: Harcourt Brace Jovanovich.
Friedman, M., and S. Kuznets. 1945. *Income from independent professional practice*. New York: National Bureau of Economic Research.
Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.
Goldberger, A. 1989. Economic and mechanical models of intergenerational transmission. *American Economic Review* 79: 504–513.
Grossman, M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80: 223–255.
Karni, E. 1973. The transactions demand for cash: Incorporation of the value of time into the inventory approach. *Journal of Political Economy* 81: 1216–1225.
Katz, L., and K. Murphy. 1992. Changes in relative wages, 1963–1987: Supply and demand factors. *Quarterly Journal of Economics* 107: 35–78.
Landes, W. 1998. Gary S. Becker Biography. In *The new Palgrave dictionary of economics and the law,* ed. P. Newman. London: Macmillan Reference.
Lazear, E. 2000. Economic imperialism. *Quarterly Journal of Economics* 115: 99–146.
Lewis, H. 1956. Hours of work and hours of leisure. *Annual Proceedings of the Industrial Relations Research Association* 12: 196–206.
Lucas Jr., R., and N. Stokey. 1983. Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics* 12: 55–93.
Mankiw, N., D. Romer, and D. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–437.
Meltzer, Allan H., and Scott F. Richard. 1981. A rational theory of the size of government. *Journal of Political Economy* 89(5): 914–927.
Mincer, Jacob. 1962. Labor force participation of married women: A study of labor supply. In *Aspects of labor economics*, ed. H.G. Lewis. Princeton: Princeton University Press.
Mulligan, Casey B. 1997. *Parental priorities and economic inequality*. Chicago: University of Chicago Press.
Mulligan, C., and Y. Rubinstein. 2005. Selection, investment, and women's relative wages since 1975. Working Paper No. 11159. Cambridge, MA: NBER.
Mulligan, C., and X. Sala-i-Martin. 1999. Gerontocracy, retirement, and social security. Working Paper No. 7117. Cambridge, MA: NBER.
Mulligan, C., R. Gil, and X. Sala-i-Martin. 2004. Do democracies have different public policies than nondemocracies? *Journal of Economic Perspectives* 18(1): 51–74.
Neal, D., and W. Johnson. 1996. The role of premarket factors in black-white wage differences. *Journal of Political Economy* 104: 869–895.
Olson, M. 1965. *The logic of collective action*. Cambridge, MA: Harvard University Press.
Peltzman, S. 1976. Toward a more general theory of regulation. *Journal of Law and Economics* 19: 211–240.
Persson, T., and G. Tabellini. 2004. Constitutions and economic policy. *Journal of Economic Perspectives* 18(1): 75–98.
Reid, M. 1952. Effect of income concept upon expenditure curves of farm families. In *Studies of Income and Wealth*, vol. 15. New York: NBER.
Robson, A., and H. Kaplan. 2003. The evolution of human life expectancy and intelligence in hunterecconomies. *American Economic Review* 93: 150–169.
Schumpeter, J. 1942. *Capitalism, socialism, and democracy*. New York: Harper and Brothers.
Shimer, R., and Ivan Werning. 2003. *Optimal unemployment insurance with sequential search*. Mimeo: University of Chicago.
Smith, J., and F. Welch. 1989. Black economic progress after Myrdal. *Journal of Economic Literature* 27: 519–564.
Summers, R., and A. Heston. 1991. The Penn world table (Mark 5): An expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106: 327–368.

Beeke, Henry (1751–1837)

S. Rashid

The Rev. Henry Beeke has hitherto been known to historians of economics for his critique of the methods by which the value of Williams Pitt's income tax had been estimated, *Observations on the Produce of the Income Tax* (1800). This pamphlet is one of the better examples of the

tradition of economics known as Political Arithmetic. It earned the praise of both J.R. McCulloch, who called it 'the best example of the successful application of statistical reasonings to finance that had then appeared', and Sir Robert Giffen, who examined the estimates with some care in *The Growth of Capital* and remarked that many of Beeke's calculations 'were fully justified by the results of the Income Tax' (p. 100).

Beeke was a good friend of such prominent Tories as Nicholas Vansittart, later Lord Addington, and J.C. Herries. It is probable that the publication of the *Observations* led to a meeting with the Younger Pitt at Addington's house in 1800. Thereafter Beeke regularly provided advice on a variety of economic topics to the Tory administration and Beeke became something of an unofficial economic adviser to the government. The topics on which he provided the most regular advice were funding and paper money. However, the most notable of Beeke's reports is one on the wheat harvest of 1800. Widespread rumours of a scarcity led Beeke to write a long report, now in the Devon Public Record Office, in which he detailed reasons why there was no real scarcity. In the process, he also provided the first clear statement of what is called a Giffen good:

> In all times of Dearness, there is an *Increase* in the consumption of whatever forms the *Basis* of the Food of the People, so long as by retrenching all other expense in Provisions they can possibly find Money to purchase it. They do not understand the Arts of Economical Cookery, they have not Utensils for it, their Stomachs are not used to novelties. With us the Consumption of Bread always increases when their Money, if divided, will not purchase an addition of Meat to the Diet which they cannot abandon. And this is true even when Bread is become in comparison far more costly.

It is not known whether this report was widely circulated, but if it was, then other early statements, such as that of the bureaucrat Simon Gray, may be indebted to Beeke.

In 1801 Beeke accepted the post of Professor of Modern History at Oxford and part of his duties involved delivering lectures on political economy, probably the earliest such lectures at Oxford. After about 1810 Beeke seems to have became less familiar with his Tory friends. In 1814 he became Dean of Bristol and for the rest of his long life appears to have eschewed all economic controversy.

See Also

- ▶ Giffen's Paradox
- ▶ Gray, Simon (Alias George Purves, LL.D.) (fl. 1795–1840)
- ▶ Political Arithmetic

Selected Works

1799. *Observations on the produce of the income-tax*. London.

References

Giffen, R. 1890. *The growth of capital*. London.
McCulloch, J.R. 1845. *The literature of political economy*. London: Longman, Brown, Green and Longmans.

Beer, Max (1864–1943)

Peter Groenewegen

Journalist and historian of socialism and economics. Born in the Polish district, Tarnobrzeg, of the then Austrian province of Western Galicia, he migrated to Germany in 1889 to work as journalist on the Volkstimme, a socialist newspaper. Subsequent political persecution, including a jail sentence, forced him to leave Germany for London in 1894. There he became one of the first students at the London School of Economics (1895–6) and until his return to Germany in 1915 he worked as London correspondent of Vorwarts. He made brief visits to Paris (1899) and New York (1900–1901). The triumph of National Socialism in 1933 caused

his second period of political exile in London, where he died in 1943.

The importance of Beer's work for economics rests on his contributions to the history of economics, two of them written during the last years of his life. His *Early British Economics* (1938) combined a pioneering study in English of medieval economics together with the more usual discussion of mercantilism. Beer's remark (1938, p. 228) that William Petty 'was in economics what Francis Bacon was in philosophy – the emancipator from Aristotle', gives some of the flavour of the work. His *Inquiry into Physiocracy* (1939) deserves praise as the first English study on the subject since Higgs's work (1897) but is a 'somewhat bizarre interpretation' based on the argument that Quesnay wished to recreate medieval economic society (Meek 1962, p. 368). Of greater interest is his still very readable and useful *History of British Socialism* (1912), particularly its detailed analysis of the sources of 'Ricardian socialism' and its consequent dichotomy between the cooperative socialism inspired by Owen and the 'economics of anti-capitalism' of Ravenstone and Hodgskin. He also wrote an excellent book on Marx Beer (1918).

Selected Works

1912. *A history of British socialism*. English edition in two volumes. London: G. Bell & Sons, 1921.
1918. *The life and teaching of Karl Marx*. Trans. from the German by T.C. Partington and H.J. Stenning. London, 1924.
1938. *Early British Economics from the XIIIth to the middle of the XVIIIth century*. London: George Allen & Unwin.
1939. *An inquiry into physiocracy*. London: George Allen & Unwin.

References

Higgs, H. 1897. *The physiocrats*. London: Macmillan.
Meek, R.L. 1962. *The economics of physiocracy*. London: George Allen & Unwin.

Beggar-Thy-Neighbour

Nilüfer Çağatay

The orthodox approach to international trade assumes full employment of a given amount of resources in the world economy. Within this framework, free trade (with certain exceptions such as the optimum tariff argument) is viewed to bring about the most efficient international division of labour, thereby maximizing world output as well as the output of individual trading economies. A corollary to this argument is that, in general, interferences with the process of free trade leave both the intervening country and its trading partners as a whole worse off compared to the free trade situation.

Joan Robinson (1937) in what is now a classic article pointed out the problematic nature of this formulation. Elaborating on Keynes's notes on mercantilism in the *General Theory,* she argued that in times of worldwide unemployment, it is, indeed, possible for one country to increase its employment and total output by increasing its trade balance at the expense of other countries. She coined the phrase 'beggar-thy-neighbour' to describe such policies.

Robinson started her argument by pointing out that an increase in the trade balance, with a given level of home investment, is equivalent to the effect of an investment increase which would normally restore the level of employment in an economy with underemployment. The change in the trade balance and subsequently in home employment, can be brought about by policies which lead to the expansion of exports and/or of import-competing production. Robinson discussed four such policies: (1) exchange rate depreciation, (2) reductions in wages, (3) subsidies to exports and (4) restrictions by means of quotas and tariffs.

According to Robinson, a fall in the exchange rate or alternatively a fall in money wages stimulates output in exporting and import-competing industries and generally increases the trade balance. Although she had pointed out that an

increase in the trade balance does not necessarily lead to higher employment, until recently, it was assumed that currency depreciation results in stimulating output and employment. It was argued that devaluation would have contractionary effects only when the Marshall–Lerner condition is violated. Since in the case of economies with underemployment this condition (that the sum of the absolute values of export and import elasticities must exceed unity) is assumed to be satisfied, currency depreciation was viewed as an output/employment stimulating device. More recently, Krugman and Taylor (1978) have discussed the contractionary effects of devaluation, pointing out that depreciation can lead to a reduction in national output if (1) imports initially exceed exports, (2) consumption propensities from wages and profits differ and (3) there are significant export taxes that cause an increase in government revenues as a result of devaluation.

There is another reason for exchange rate depreciation and/or reduction in money wages not to act as output stimulating policies. These two policies, if they succeed at all, stimulate exports and reduce imports if international competition takes place through prices. However, there are significant non-price factors in international competition which limit the role of devaluation and money wage reductions in increasing the trade balance or restoring full employment.

Historically, import controls and export subsidies have proved to be more effective devices. Import controls by means of tariffs and quotas are expected to increase the trade balance by protecting import competing sectors from external competition. Subsidies to exports are argued to increase the international competitiveness of such sectors.

For any of these expedients to succeed in terms of increasing employment in an economy, it is necessary that its trade partners do not retaliate. However, as Robinson pointed out, in times of general unemployment, a nation increasing its trade balance is faced with retaliation by others. What begins as a beggar-thy-neighbour remedy for unemployment in one country turns into an international beggar-thy-neighbour game with the total volume of international trade shrinking relative to world output and eventually leading to a decline in world economic output. Indeed, it was this kind of competitive behaviour that characterized the international trade policies of the thirties.

The postwar structuring of the world trade system through the establishment of institutions such as the IMF and GATT was based on efforts to avoid beggar-thy-neighbour policies. However, the new system failed to 'frame rules that would permit the right exceptions while ruling out the wrong ones'; it did not establish when the very same policies would be bad-neighbourly and when they would be good-neighbourly (Robinson 1965). The rules that prohibit the use of devices that boost exports and check imports also prevent individual countries from employing them in times of necessity in constructive ways from the point of the world economy as a whole. For instance, in the case of a country that is attempting to stimulate its output by increasing its investment, initially imports may rise faster than exports. Such an economy would need to reduce its propensity to import while keeping its total level of imports constant so as not to develop a balance of payments problem.

The new system, instead of allowing the necessary exceptions, institutionalized a different type of bad-neighbourly conduct by advocating deflationary policies as the remedies for the balance of payments problems. In fact, institutions such as the IMF became the executioners of these deflationary policies which aim to bring about balanced trade starting from a deficit position by inducing a slump to cut down imports.

In recent times, these issues and problems have re-emerged under the present situation of the world economy in which there is widespread unemployment and the pressure for individual countries to respond with beggar-thy-neighbour policies has been building. The arguments developed by Joan Robinson have a great deal of freshness today in the light of these circumstances, making her contributions as relevant as when she first formulated them in the 1930s.

See Also

► Robinson, Joan Violet (1903–1983)

References

Krugman, P., and L. Taylor. 1978. Contractionary effects of devaluation. *Journal of International Economics* 8(3): 445–456.

Robinson, J. 1937. Beggar-my-neighbour remedies for unemployment. In *Essays in the theory of unemployment*, ed. J. Robinson. London: Macmillan.

Robinson, J. 1965. The new mercantilism. An inaugural lecture delivered at the University of Cambridge. In *Collected economic papers of Joan Robinson*, vol. 4. Oxford: Basil Blackwell, 1973.

Behavioural Economics

Herbert A. Simon

Since economics is certainly concerned with human behaviour – with, as Marshall put it, '[the] study of mankind in the ordinary business of life' – the phrase 'behavioural economics' appears to be a pleonasm. What non-behavioural economics can we contrast with it? The answer to this question is found in the specific assumptions about human behaviour that are made in neoclassical economic theory.

Contrast of Behavioural with Neoclassical Theory

The neoclassical assumptions. How does human behaviour enter into classical and neoclassical economics? First, human goals and motivations are assumed to be given a priori in the form of a utility function, which allows an individual to make consistent choices among all possible bundles of goods and services. Second, economic actors are assumed always to choose, among the alternatives open to them, that one of the alternatives that yields the greatest utility (Savage 1954).

These two assumptions – of a given utility function and of utility maximization (rationality) – are usually made explicitly. Other assumptions about human behaviour are often implicit in classical and neoclassical theory, and are not necessarily maintained through all variants of the theory. It is usually assumed that not only the utility function but also the set of available alternatives is given a priori. In search theory, this assumption is replaced by the assumption that new alternatives may be generated by a process of search, but at some cost, which is assumed to be known, as is the expected marginal return to the search.

With respect to the consequences of alternatives, it may be assumed that these are known completely and with certainty, or that what is known is a joint probability distribution of outcomes, although occasionally, forms of 'uncertainty' that are not reducible to probabilities are introduced into the theory. It is almost always assumed in neoclassical theory that, given their knowledge of utilities, alternatives, and outcomes, economic actors can compute which alternative will yield the greatest (expected) utility – although it is conceptually (if seldom practically) possible to incorporate a cost of computation into the theories that is analogous to the cost of generating alternatives in search theory.

Behavioural departures from neoclassical assumptions. With this characterization of classical and neoclassical economics, we can now, by a process of contrast, define rough boundaries for behavioural economics. Behavioural economics is concerned with the empirical validity of these neoclassical assumptions about human behaviour and, where they prove invalid, with discovering the empirical laws that describe behaviour correctly and as accurately as possible. As a second item on its agenda, behavioural economics is concerned with drawing out the implications, for the operation of the economic system and its institutions and for the public policy, of departures of actual behaviour from the neoclassical assumptions. A third item on its agenda is to supply empirical evidence about the shape and content of the utility function (or of whatever construct will replace it in a empirically valid behavioural theory) so as to strengthen the predictions that can be made about human economic behaviour.

Thus, behavioural economics is best characterized not as a single specific theory but as a commitment to empirical testing of the neoclassical assumptions of human behaviour and to

modifying economic theory on the basis of what is found in the testing process. And not all of the economists who hold a behavioural point of view also hold a common theory, or are all preoccupied with examining the same parts of the economic mechanism.

Directions of behavioural research. Accordingly, we can distinguish a number of different foci and directions of inquiry in behavioural economics. Some investigations are concerned with the assumptions of utility and profit maximization, and with replacing these with alternative motivational assumptions that appear to describe human motivations in the marketplace more accurately (e.g. Baumol 1959).

Another focus of behavioural research in economics is decision making under uncertainty – determining whether economic actors are able to and do maximize subjective expected utility, as called for by neoclassical theory. Here the interest is less in motivation than in the *ability* of human beings to carry out the calculations required to make the optimal decisions – the issues involved are largely cognitive.

The limitation in human ability to deal with uncertainty is just a special case of the numerous cognitive limitations that prevent economic actors from knowing and adopting the optimizing alternative of choice. The term 'bounded rationality' has been proposed to denote the whole range of limitations on human knowledge and human computation that prevent economic actors in the real world from behaving in ways that approximate the predictions of classical and neoclassical theory: including the absence of a complete and consistent utility function for ordering all possible choices, inability to generate more than a small fraction of the potentially relevant alternatives, and inability to foresee the consequences of choosing alternatives, including inability to assign consistent and realistic probabilities to uncertain future events (Simon 1955).

Conventional Behaviour

At the farthest remove from neoclassical economics are explanations of phenomena that do not rest at all on rationality assumptions. For example, it has been observed that the mean compensation of the top executive in corporations varies with the logarithm of the size of the corporation (Roberts 1959). To explain this regularity in neoclassical terms one has to show that the marginal contribution of the top executive is proportional to the logarithm of company size; and this proposition implies, in turn, very specific conditions on the distribution of executive abilities (Lucas 1978).

However, an explanation that requires no assumption about executive abilities can be derived from (a) the empirical observation that most companies are pyramidal, so that the number of organizational levels grows logarithmically with the number of employees; and (b) the empirical observation that most people regard it as 'legitimate' or 'appropriate' for a boss to be paid some multiple (about 1.5 times, say) of the salary of his or her immediate subordinates. The observed regularity in average salaries follows from these two observations (Simon 1957). We may call this a 'sociological' explanation, since it postulates commonly held social beliefs or attitudes, but not rational calculation.

In the same way, the observed regularities in business firm size distributions (which usually fit closely to the Pareto distribution) follow from the assumption that expected growth is proportional to attained size (the Gibrat assumption). This assumption, in turn, can be derived from the postulate that access to internal and external investment funds is proportional to size, without postulating rational choice as part of the causal mechanism (Ijiri and Simon 1977).

As a third example, the empirical observation that the labour share of total product has been nearly constant in the industrialized world during the past century may be explained from the premises: (a) that interest rates are nearly stable; (b) that at all levels of average per capita income, nearly the same fraction of total income is saved, hence the ratio of capital supply to total output is nearly constant (Simon 1979). But premises (a) and (b) may be accepted as empirical (sociological) regularities that do not derive from assumptions of rationality or from assumptions that marginal costs are equal to marginal benefits.

If, in fact, important social phenomena, like salaries, access to capital for growth, or rates of saving, do not depend on rational calculations, but are conventionally determined, then the corresponding parts of economic theory need to be built on empirical knowledge of socially accepted conventions rather than on derivations from the assumption of rationality. What phenomena are 'conventionally' rather than 'rationally' determined is itself an empirical question, and not one that can be settled by pure reasoning.

It is often possible to rationalize the kinds of behaviour that were described above as conventional, but the real work in such explanation is done by ad hoc auxiliary assumptions, which is quite different from inferring the behaviour uniquely from the assumptions of economic rationality alone. For example, almost any observed distribution of executive salaries would be compatible with Lucas's (1978) model provided that appropriate adjustments were made in the assumed (and unobservable) distribution of executive abilities. The proposed sociological explanation for the salary distribution is falsifiable, while Lucas's model is not.

In similar fashion, the historical stability of interest rates and of the saving to income ratio could be attributed in some measure to characteristics of individual utility functions. In this case, again, rationality assumptions play no important role. The argument rests on an unmotivated assumption about human preferences – an empirical assumption.

The 'New Institutional Economics'

Less distant from neoclassical theory than the examples just cited are explanations that incorporate the rationality assumptions, but also invoke limits on the information available to actors or impose 'transaction costs' on their use of information in order to account for specific institutional phenomena. Much of the work of Williamson (1975) falls in this category.

For example, there is a fundamental difference between a sales contract and an employment contract. The former involves an exchange of specific commodities for money, while the latter involves the willingness of the employee to accept authority (i.e. to have his actions determined by the employer) in exchange for money. How can we predict which economic transactions will take the form of sales contracts, and which the form of employment contracts? It can be argued that if the employer has great uncertainty as to what specific duties he will want performed, and the employee is nearly indifferent among various ways of spending his time, then an employment contract will be the rational choice, otherwise a sales contract (Williamson 1975, chs. 4, 5; Simon 1951).

Similar rational analyses have been proposed to explain the existence of various kinds of contractual instruments – for example, special forms of insurance and of forward contracting. 'Moral risk', that is, the practical unenforceability of some kind of contract clauses in the face of opportunistic behaviour of the parties, also commonly enters into explanations for the existence or non-existence of particular sorts of contracts. Williamson (1975) invokes the combination of bounded rationality and opportunism as the mechanism for explaining the relative roles of banks, conglomerates and divisionalized corporations in allocating investment capital.

The body of modern economic analysis that employs concepts like limited information, transaction costs, and opportunism to explain observed economic phenomena is often called the 'New Institutional Economics'. A common feature of these sorts of institutional analyses is that the real 'action' in the derivations comes not from the rationality assumptions, but from the assumptions of informational or other limits on rationality, or from what were in the previous section called sociological postulates. If employers had perfect foresight, or if they could costlessly renegotiate their contracts with employees for each new task to be performed, there would be no rationale for employment contracts. If householders made even roughly correct estimates of flood risk, purchases of flood insurance would rise sharply. Conglomerates can allocate investment funds more profitably than banks, because the former have inside information not available to the latter. And so on.

Moreover, since the institutional analyses usually involve the comparison of two, or a few, discrete alternatives rather than a continuum of choices, it is seldom necessary to evoke the assumption of maximization. Even a satisficing actor can be expected to select the better of two alternatives if the difference in expected outcomes is large. Conversely, if institutional arrangements are compatible with either maximizing or satisficing assumptions, evidence about them cannot be used to choose between these assumptions.

Explanations of institutional arrangements can often be reached by qualitative arguments about what is 'functional' rather than arguments about what is optimal. This reduction in dependence on rationality assumptions cuts two ways. By weakening the assumptions required for the arguments, it raises the prior probability of the explanation; by introducing into consideration a host of different potential auxiliary sociological assumptions, it increases the urgency of testing independently the empirical validity of these assumptions. Casual empiricism is especially inappropriate in these contexts.

It would be a valuable exercise to determine what part of neoclassical economics, and especially the new institutional economics, would survive the replacement of optimizing by satisficing or functional arguments. Presumably, theorems about Pareto optimality and market efficiency would be lost, but not necessarily theorems about stability of equilibrium. And as suggested above, many claims about the functionality of particular institutional arrangements would still be supported by the weaker assumptions.

The 'Utility Function' of the Firm

A good deal of neoclassical economic reasoning does not require any specific assumption about the shapes of the actors' utility functions. Much of the literature on the theory of the firm, however, requires the assumption that firms maximize profit or, in long-term analyses, the present value of stockholders' equity. There are many conceivable alternatives to the profit-maximization assumption, classifiable into three main categories: (a) that the firm seeks to maximize some other quantity than profit; (b) that individual executives strive to maximize their personal utilities, which are unlikely to coincide with the firm's utility; and (c) that executives and other participants identify with the subgoals of the organizational units to which they belong, and seek to maximize attainments of these subgoals (Marris 1964).

An example of the first kind of alternative is Baumol's (1959) proposal that firms strive to maximize revenue rather than profits. Evolutionary arguments have been evoked against this sort of alternative to profit maximizing, but the objections rest on the assumption, much stronger than any in biological Darwinism, that only profit maximizers can survive. Again, it is clear that the issue has to be decided by empirical inquiry. In the biological world at least, many organisms survive that are not maximizers but that operate at far less than the highest achievable efficiency. Their survival is not threatened as long as no other organisms have evolved that can challenge the possession of their specific niches. Analogously, since there is no reason to suppose that every business firm is challenged by an optimally efficient competitor, survival only requires meeting the competition. In a system in which there are innumerable rents, of long-term and short-term duration, even egregious sub-optimality may permit survival. Nelson and Winter (1982) have examined evolutionary models of business firm growth that dispense with the assumption of profit maximization. Each firm, in their models, may have a different production function, and one that changes through imitation or, stochastically, from investment in R&D. With industry models of this sort, distributions of firm sizes may be obtained similar to those actually observed.

The second kind of deviation from the profit maximizing assumption – that which rests on considerations of executive behaviour – is perhaps the most frequently advanced. One form it takes is the 'organizational slack' hypothesis of Cyert and March (1963), which suggests that firms will ordinarily settle for 'satisfactory' profits, and that it is only when they fail to achieve these that they search for improved products or methods of operation. Leibenstein's (1976)

concept of 'X-efficiency' has a similar flavour. One way to relate such behaviour to rationality assumptions is to postulate that 'comfort' is an important component in executive utility functions, and that there is a trade-off between the comfortable executive life and profits.

Another deviation from the neoclassical assumptions is to view the firm not as a system with a well-defined utility function, but as a coalition of partially cooperating and partially competing interests. The goal-defining process can then be viewed in terms of game theory. The notion of the corporation as a coalition has been developed by Cyert and March (1963) and by Williamson (1975).

Research on organizations has revealed the central importance of the mechanism of *identification* – the tendency of actors in an organization to internalize, and be guided by, the goals of the particular subparts of the organization with which they are most closely associated. In part, identification may be accounted for by systems of reward, but that is almost certainly an incomplete account. Cognitive mechanisms (especially focus of attention) and mechanisms of social motivation (including docility) appear to be at least as important as rewards in determining the criteria of choice that are applied to decisions. For example, departmental executives asked to identify the most important problem facing a company that has been described to them tend disproportionately to select problems in their own domain of expertise – the sales managers, sales problems; the production managers, production problems, and so on (Dearborn and Simon 1958).

Individual Utility

In analysing the behaviour of consumers and employees, the outcomes often depend heavily on what is assumed about the structure and content of their utility functions. Changes in preferences between work and leisure, for example, will ordinarily produce corresponding changes in the labour supply function. Similarly, changes in social attitudes about women's roles may play a major part in determining the participation of women in the labour force, the number of children, and many other variables important to long-term social and economic development. As Becker (1981) has shown, this does not forestall economic analysis of these phenomena, but it does limit the power of the rationality principle to arrive at conclusions without numerous assumptions about the utility function.

The assumption of utility maximization is sometimes misunderstood to imply that only selfish motives play a role in human behaviour. However, except in the context of evolutionary theories that emphasize selection, the assumption of utility maximization, whether of business firms or of individuals, does not, of course, imply that human beings are selfish, only that they are consistent. Altruism can be accommodated in the utility function simply by including the well-being of other persons as one of its components. When utility maximization is not postulated directly, but is derived from Darwinian arguments of survival, then altruism, except for 'weak altruism' or 'enlightened self-interest', becomes more difficult to account for. Maximization of Darwinian fitness, as usually formulated, leaves little room for altruistic behaviour towards others who are not close relatives.

It is possible that this difficulty could be removed by closer attention to 'docility' (susceptibility to social instruction and influence) as a trait contributing strongly to fitness (Simon 1983, ch. 2). Whether docility does, in fact, play a role in shaping preferences is an empirical question to be answered by behavioural research. And to what extent altruism (as distinguished from enlightened self interest) actually enters into human behaviour is also an issue that must be settled by empirical study.

Decisions Under Uncertainty

Of all of the variables affecting, potentially or actually, the economic decision-making process, uncertainty has perhaps attracted the greatest amount of research attention, both theoretical and empirical. We now have substantial knowledge about the relation between actual human

behaviour in the face of uncertainty and the behaviour predicted by the subjective expected utility model.

Data from the laboratory shows that, under different circumstances of choice, subjects depart from the predictions of the SEU model in diametrically opposite directions. At times when they view the world as stable or static, they place too much weight on past events in prediction; but when they perceive large structural changes taking place in the environment, they underestimate the significance of past experience for predicting the future.

Extensive studies of decisions to purchase or not to purchase flood insurance (Kunreuther 1978) reveal behaviour that cannot be reconciled with any model of utility maximization. Specifically, it is found that people tend to ignore (hence, not to insure against) low-probability, high-consequence events, unless they have had rather direct past personal experience of them.

Assumptions about how people deal with uncertainty and predict future events have important consequences for macroeconomic theory. For example, rational expectations theories, adaptive expectations theories and cobweb theories make quite different predictions about the impact of government monetary and fiscal policies. But it can be shown that the bounds placed on rationality, and not the rationality assumptions, account for the main differences in prediction among, for example, leading business-cycle theories. A Keynesian theory of the cycle can be derived from a neoclassical model simply by assuming that labourers suffer from money illusion in their demands for wages; while a rational-expectations theory of the cycle can be derived from the same neoclassical model by assuming that businessmen mistake a general change in the price level for a relative change in the prices of the goods they purchase (another form of money illusion). As in institutional arguments, the real work in these theories is not being done by the rationality assumptions but by auxiliary assumptions about limits on rationality (including limits on the accuracy of information) – assumptions often made on the basis of extremely casual evidence.

Search and Choice Processes

Applications of the concept of bounded rationality to choice situations where uncertainty about outcomes is a dominant factor have been discussed above. Another line of research, mainly in the laboratory, seeks to examine the processes of search that people use, both in the exploration for alternatives and in their use of information about alternatives (Hogarth 1980; Kahneman et al. 1982). A standard experimental paradigm confronts subjects with a number of alternatives, and allows them to obtain additional information about each until they make a choice or the available information is exhausted. Experiments in this paradigm show that decision makers usually satisfice, both in the sense of failing to examine all of the information that is available, and in the sense of choosing an alternative as soon as one has been found that is satisfactory along all the dimensions of concern.

A few studies have been made of situations where subjects must explore for new alternatives and must decide when to terminate the search. For example, a field study of the job search processes of business school students (Soelberg 1966) showed students using a variety of rules of thumb to limit the list of firms they contacted and to choose among those who made offers to them.

In real-world situations, it is seldom realistic to talk about examining all alternatives or paying attention to all the potentially relevant information. Empirical evidence is still very scanty about the circumstances under which people will pay attention to particular variables in making their decisions (e.g. under what circumstances they will pay attention to the difference between real and nominal prices). Evidence is also very scanty (but see Cyert and March 1963; Soelberg 1966) about the circumstances under which people search for new alternatives.

The search for alternatives is, of course, a critical process in understanding entrepreneurial behaviour, decisions to invest in research and development activities, and generally in understanding the ways in which new economic activities and enterprises are spawned (Nelson and

Winter 1982; Winter 1984). This Schumpeterian view toward economic development has recently been expanded and systematized by Nelson and Winter (1982), who model the growth of industries in which the development of new products and practices plays a major role in competition among firms. They show that in such models, evolutionary selection can replace or complement optimization as a driving force for change.

Methods of Behavioural Research

Within the classical tradition, the principal evidence that has been used to test economic theories empirically has been public statistics, usually aggregated at least to the level of the industry (although limited use has been made, especially in investment theory, of financial data for individual firms). A powerful and sophisticated set of econometric tools has been developed for extracting from such data all the information that they contain. But it becomes increasingly clear that data of these kinds are simply too aggregated and noisy to reveal much about the decision-making processes of the economic actors. In neoclassical theory, those processes are simply postulated, in terms of the rationality assumptions, and never subjected to any really searching direct test.

What is even more troublesome, the typical inputs to econometric analysis provide little information that would be relevant to choosing the appropriate auxiliary assumptions (assumptions about the limits to rationality) that, we have seen, play a major role in drawing inferences from economic theories. In the absence of empirical evidence for choosing these assumptions, they are generally made in a casually empirical, armchair way.

The progress of economics, and especially the prospects for adequate empirical testing of economic theories, would seem to depend, therefore, on finding new kinds of data to supplement the sorts of aggregative evidence now typically employed. One important new kind of data comes from case studies, a second from survey research, and a third from laboratory experiments. Computer simulation models can provide a powerful tool for relating these kinds of data to theory. But successful use of such data will call for new methods for aggregating data that are gathered from individual firms or individual economic actors. Each of these points calls for a brief comment.

Case studies. A small, but growing, number of case studies have been made of the decision-making process in individual business firms (Cyert and March 1963; Bromiley 1981; etc.) and of individuals (Soelberg 1966; Clarkson 1962; Bouwman 1982; etc.). That these studies have had little impact on economic theory must be attributed in considerable measure to the absence of a theory of aggregation that would indicate just how to use them. In general, economists, though willing to engage in casual introspective empiricism, have not been willing to treat the firms whose behaviour has been studied in depth as 'representative firms'.

Survey research. In the case of survey research, where appropriate sampling methods can be used, there is no such difficulty in relating the survey data to macromodels of industries or the economy. In fact, data on businessmen's expectations have been used, to a limited extent, as inputs into econometric models (Katona 1975). The main limiting factors here appear to be the small number of economists who are trained to produce and interpret survey data, and the limited resources that have been applied to generating such data.

Experiments. The use of experiments to study economic behaviour is a relatively new, and rather rapidly spreading, development (see, e.g., Smith 1976; Hong and Plott 1982). A principal problem here is to produce in the laboratory motivational conditions that can be extrapolated to the real world. A principal limitation on the growth of laboratory experimentation is, again, the limited access of graduate students in economics to training in techniques of experimentation.

Computer modelling. Computers are widely used in economics, not only to run regressions, but also to model the economic system. Most of the models are aggregative, but there has been a certain amount of investigation of so-called 'micro-models', whose units are samples of individual actors and firms (e.g. Eliasson 1984; Winter 1984). In addition, one can point to a few

models of decision-making within an individual firm (Bonini 1963; Bromiley 1981). Attention has been paid to the aggregation problem in constructing micro-models, but that problem remains a major barrier to acceptance of findings derived from models of these sorts as a part of the main stream of economic theory and knowledge.

See Also

- ▶ Rationality, Bounded
- ▶ Satisficing

Bibliography

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

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

Bonini, C.P. 1963. *Simulation of information and decision systems in the firm*. Englewood Cliffs: Prentice-Hall.

Bouwman, M.J. 1982. The use of accounting information: Expert versus novice behavior. In *Decision making*, ed. Ungson and Braunstein. Boston: Kent Publishing Co.

Clarkson, G.P.E. 1962. *Portfolio selection: A simulation of trust investment*. Englewood Cliffs: Prentice-Hall.

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

Dearborn, D.C., and H.A. Simon. 1958. Selective perception. *Sociometry* 21: 140–144.

Eliasson, G. 1984. Micro heterogeneity of firms and the stability of industrial growth. *Journal of Economic Behaviour and Organization* 5: 249–274.

Hogarth, R.M. 1980. *Judgment and choice: The psychology of decision*. New York: Wiley.

Hong, J.T., and C.R. Plott. 1982. Rate filing policies for inland water transportation: An experimental approach. *Bell Journal of Economics* 13: 1–19.

Ijiri, Y., and H.A. Simon. 1977. *Skew distributions and the sizes of business firms*. Amsterdam: North-Holland.

Kahneman, D., P. Slovic, and A. Tversky (eds.). 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Katona, G. 1975. *Psychological economics*. New York: Elsevier Publishing Co.

Kornai, J. 1971. *Anti-equilibrium*. Amsterdam: North-Holland.

Kunreuther, H., et al. 1978. *Disaster insurance protection: Public policy lessons*. New York: Wiley.

Leibenstein, H. 1976. *Beyond economic man*. Cambridge, MA: Harvard University Press.

Lucas Jr., R.E. 1978. On the size distribution of business firms. *Bell Journal of Economics* 9: 508–523.

Marris, R. 1964. *The economic theory of 'managerial' capitalism*. New York: Macmillan.

Nelson, R.R., and S.G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Harvard University Press.

Roberts, D.R. 1959. *Executive compensation*. Glencoe: The Free Press.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Simon, H.A. 1951. A formal theory of the employment relationship. *Econometrica* 19: 293–305. Reprinted as ch. 5.2 in Simon (1982).

Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69: 99–118. Reprinted as ch. 7.2 in Simon (1982).

Simon, H.A. 1957. The compensation of executives. *Sociometry* 20: 32–35. Reprinted as ch. 5.6 in Simon (1982).

Simon, H.A. 1979. On parsimonious explanations of production relations. *Scandinavian Journal of Economics* 81: 459–474. Reprinted as ch. 4.4 in Simon (1982).

Simon, H.A. 1982. *Models of bounded rationality*, 2 vols. Cambridge, MA: MIT Press.

Simon, H.A. 1983. *Reason in human affairs*. Stanford: Stanford University Press.

Smith, V.L. 1976. Experimental economics: Induced value theory. *American Economic Review* 66: 274–279.

Soelberg, P. 1966. *A study of decision making: Job choice*. Cambridge, MA: Alfred P. Sloan School of Management, MIT.

Williamson, O.E. 1975. *Markets and hierarchies*. New York: Free Press.

Winter, S.G. 1984. Schumpeterian competition in alternative technological regimes. *Journal of Economic Behavior and Organization* 5: 287–320.

Behavioural Economics and Game Theory

Faruk Gul

Abstract

Behavioural economics, broadly defined, refers to the research programme that investigates the relationship between psychology and economic behaviour. The purpose of this article is to provide an outline of behavioural economics research and to describe where research in behavioural game theory stands within this outline. The aim is not to assess the impact of particular contributions or

describe and interpret specific applications. Rather, the goal is to provide an organization of the literature based on the type of departures from standard theory.

Keywords

Ambiguity; Ambiguity aversion; Asymmetric information; Auctions; Behavioural economics; Behavioural economics and game theory; Commitment; Evolutionary games; Expected utility hypothesis; Fixed point theorems; Framing effects; Game theory; Independence axiom; Interdependent preferences; Mechanism design; Neuroeconomics; Preference reversals; Probability distributions; Prospect theory; Psychological games; Social preferences; Ultimatum game; Uncertainty

JEL Classifications

C7

In traditional economic analysis, as well as in much of behavioural economics, the individual's motivations are summarized by a utility function (or a preference relation) over possible payoff-relevant outcomes while his cognitive limitations are described as incomplete information. Thus, the standard economic theory of the individual is couched in the language of constrained maximization and statistical inference.

The approach gains its power from the concise specification of payoff-relevant outcomes and payoffs as well as a host of auxiliary assumptions. For example, it is typically assumed that the individual's preferences are well behaved: that is, they can be represented by a function that satisfies conditions appropriate for the particular context such as continuity, monotonicity, quasi-concavity, and so on. When studying behaviour under uncertainty, it is often assumed that the individual's preference obeys the expected utility hypothesis. More importantly, it is assumed that the individual's subjective assessments of the underlying uncertainty are reasonably close to the observed distributions of the corresponding variables. Even after all these bold assumptions, the standard model would say little if the only relevant observation regarding the utility function is one particular choice outcome. Thus, economists will often assume that the same utility function is relevant for the individual's choices over some stretch of time during which a number of related choices are made. One hopes that these observations will generate enough variation to identify the decision-maker's (DM's) utility function. If not, the analyst may choose to utilize choice observations from different contexts to identify the individual's preferences or make parametric assumptions. The analyst may even pool information derived from observed choices of different individuals to arrive at a representative utility function.

Experimental Challenges to the Main Axioms of Choice Theory

The simplest type of criticism of the standard theory accepts the usual economic abstractions and the standard framework but questions specific assumptions within this framework.

The Independence Axiom

Allais (1953) offers one of the earliest critiques of standard decision-theoretic assumptions. In his experiment, he provides two pairs of binary choices and shows that many subjects violate the expected utility hypothesis, in particular, the independence axiom. Allais's approach differs from the earlier criticisms: Allais questions an explicit axiom of choice theory rather than a perceived implicit assumption such as 'rationality'. Furthermore, he does so by providing a simple and clear experimental test of the particular assumption.

Subsequent research documents related violations of the independence axiom and classifies them. Researchers have responded to Allais's critique by developing a class of models that either abandons the independence axiom or replaces it with weaker alternatives. The agents in these models still maximize their preference and still reduce uncertainty to probabilistic assessments (that is, they are probabilistically sophisticated), but have preferences over lotteries that fail the independence axiom.

Non-expected utility preferences pose a difficulty for game theory: because many non-expected utility theories do not lead to quasi-concave utility functions, standard fixed point theorems cannot be used to establish the existence of Nash equilibrium. Crawford (1990) shows that if one interprets mixed strategies not as random behaviour but as the opponents' uncertainty regarding this behaviour, then the required convex-valuedness of the best response correspondence can be restored and existence of Nash equilibrium can be ensured.

In dynamic games, abandoning the independence axiom poses even more difficult problems. Without the independence axiom, conditional preferences at a given node of an extensive form game (or a decision-tree) depend on the unrealized payoffs earlier in the game. The literature has dealt with this problem in two ways: first, by assuming that the DM maximizes his conditional preference at each node (for a statement and defence of this approach, see Machina 1989). This approach leads to dynamically consistent behaviour, since the DM ends up choosing the optimal strategy for the reduced (normal form) game. However, it is difficult to compute optimal strategies once conditional preference depends on the entire history of unrealized outcomes. The second approach rejects dynamic consistency and assumes that at each node the DM maximizes his unconditional preference given his prediction of future behaviour. Thus, in the second approach, each node is treated as a distinct player and a subgame perfect equilibrium of the extensive form game is computed. Game-theoretic models that abandon the independence axiom have favoured the second approach. Such models have been used to study auctions.

Redefining Payoffs: Altruism and Fairness

The next set of behavioural criticisms question common assumptions regarding deterministic outcomes. Consider the *ultimatum game:* Player 1 chooses some amount $x \leq 100$ to offer to Player 2. If Player 2 accepts the offer, 2 receives x and 1 receives $100 - x$; If 2 rejects, both players receive 0. Suppose the rewards are measured in dollars and Player 1 has to make his offer in multiples of a dollar. It is easy to verify that if the players care only about their own financial outcome, there is no subgame perfect Nash equilibrium of this game in which Player 1 chooses $x > 1$. Moreover, in every equilibrium, any offer $x > 0$ must be accepted with probability 1. Contrary to these predictions, experimental evidence indicates that small offers are often rejected. Hence, subjects in the Player 2 role resent either the unfairness of the (99,1) outcome, or Player 1's lack of generosity. Moreover, many experimental subjects anticipate this response and make more generous offers to ensure acceptance. Even in the version of this game in which Player 2 does not have the opportunity to reject (that is, Player 1 is a *dictator*), Player 1 often acts altruistically and gives a significant share to Player 2.

More generally, there is empirical evidence that suggests that economic agents care not only about their physical outcomes but also about the outcomes of their opponents and how the two compare. Within game theory, this particular behavioural critique has been influential and has led to a significant theoretical literature on social preferences (see, for example, Fehr and Schimdt 1999).

Redefining the Objects of Choice: Ambiguity, Timing of Resolution of Uncertainty, and Preference for Commitment

The next set of behavioural criticisms points out how the standard definition of outcome or consequence is inadequate. The literature on ambiguity questions probabilistic sophistication; that is, the idea that all uncertainty can be reduced to probability distributions. Ellsberg (1961) provides the original statement of this criticism. Consider the following choice problem: there are two urns; the first contains 50 red balls and 50 blue balls; the second contains 100 balls, each of which is either red or blue. The DM must select an urn and announce a colour. Then a ball will be drawn from the urn he selects. If the colour of the ball is the same as the colour the DM announces, he wins 100 dollars. Otherwise the DM gets zero. Experimental results indicate that many DMs are indifferent between (urn 1, red) and (urn 1, blue) but they strictly prefer either of these choices to (urn

2, red) and (urn 2, blue). If the DM were probabilistically sophisticated and assigned probability p to choosing a red ball from urn 1 and q to choosing a red ball from urn 2, the preferences above would indicate that $p = 1 - p$, $p > q$, and $p > 1 - q$, a contradiction. Hence, many DMs are not probabilistically sophisticated.

Ellsberg's experiment has lead to choice-theoretic models where agents are not probabilistically sophisticated and have an aversion to ambiguity; that is, the type of uncertainty associated with urn 2. Recent contributions have investigated auctions with ambiguity-averse bidders and mechanism design with ambiguity aversion.

Other developments in behavioural choice theory that fall into this category have had limited impact on game-theoretic research. For example, Kreps and Porteus (1978) introduce the notion of a temporal lottery to analyse economic agents' preference over the timing of resolution of uncertainty. The Kreps–Porteus model has been extremely influential in dynamic choice theory and asset pricing but has had less impact in strategic analysis.

Kreps (1979) takes as his primitive individuals' preferences over sets of objects. Hence, an object similar to the indirect utility function of demand theory defines the individual. Kreps uses this framework to analyse preference for flexibility. So far, there has been limited analysis of preference for flexibility in strategic problems.

Gul and Pesendorfer (2001) use preferences over sets to analyse agents who have a preference for commitment (an alternative approach to preference for commitment is discussed in section "Preference Reversals"). The GP model has been used to analyse some mechanism design problems.

Limitations of the Decision-Maker

The work discussed in section "Experimental Challenges to the Main Axioms of Choice Theory" explores alternative formulations of economic consequences to identify preference-relevant considerations that are ignored in standard economic analysis. The work discussed in this section provides a more fundamental challenge to standard economics. This research seeks alternatives to common assumptions regarding economic agents' understanding of their environments and their cognative/computational abilities.

Biases and Heuristics

Many economic models are stated in subjectivist language. Hence probabilities, whether they represent the likelihood of future events or the individual's own ignorance of past events, are the DMs' personal beliefs rather than objective frequencies. Similarly, the DM's utility function is a description of his behaviour in a variety of contingencies rather than an assessment of the intrinsic value of the possible outcomes. Nevertheless, when economists use these models to analyse particular problems, the subjective probabilities (and sometimes other parameters) are often calibrated or estimated by measuring objective frequencies (or other objective variables).

Psychology and economics research has questioned the validity of this approach. Tversky and Kahnemann (1974) identify systematic biases in how individuals make choices under uncertainty. This research has led to an extensive literature on heuristics and biases. Consider the following:

(a) Which number is larger $P(A|B)$ or $P(A \cap C|B)$? Clearly, $P(A|B)$is the larger quantity; conditional on B or unconditionally, $A \cap C$ can never be more likely than A. Yet, when belonging to set C is considered 'typical' for a member of B, many subjects state that $A \cap C$ conditional on B is more likely than A conditional on B.
(b) Randomly selected subjects are tested for a particular condition. In the population, 95 per cent are healthy. The test is 90 per cent accurate; that is, a healthy subject tests negative and a subject having the condition tests positive with probability 0.9. If a randomly chosen person tests positive, what is the probability that he is ill? In such problems, subjects tend to ignore the low prior probability of having the condition and come up with larger estimates than the correct answer (less than one-third in this example).

Eyster and Rabin's (2005) analysis of auctions offers an example of a strategic model of biased

decision-making. This work focuses on DMs' tendency to overemphasize their own (private) information at the expense of the information that is revealed through the strategic interaction.

Evolution and Learning

As in decision theory, it is possible to state nearly all the assumptions of game theory in subjectivist language (see, for example, Aumann and Brandenburger 1995).

Hence, one can define Nash equilibrium as a property of players' beliefs. Of course, Nash equilibrium beliefs (together with utility maximization) will impose restrictions on observable behaviour, but these restrictions will fall short of demanding that the observed frequency of actions profiles constitute a Nash equilibrium. The theory of evolutionary games searches for dynamic mechanisms that lead to equilibrium behaviour, where equilibrium is identified with observable decisions (as opposed to beliefs) of individuals. The objective is to describe how equilibrium may emerge and which equilibria are more likely to emerge through repeated interaction in a setting where the typical epistemic assumptions of equilibrium analysis fail initially. Thus, such models are used both to justify Nash (or weaker) equilibrium notions and to justify refinements of these notions.

Cognitive Limitations and Game Theory

Some game theoretic solution concepts require iterative procedures. For example, computing rationalizable outcomes in normal form games or finding backward induction solutions in extensive form games involves an iterative procedure that yields a smaller game after each step. The process ends when the final game, which consists exclusively of actions that constitute the desired solution, is reached. In principle, the number of steps needed to reach the solution can be arbitrarily large. Ho et al. (1998) observe that experimental subjects appear to carry out at most the first two steps of these procedures.

This line of work focuses both on organizing observed violations of standard game theoretic solutions concepts and interpreting the empirical regularities as the foundation of a behavioural notion of equilibrium.

Alternative Models of the Individual

The work discussed in this section poses the most fundamental challenge to the standard economic model of the individual. This work questions the usefulness of constrained maximization as a framework of economic analysis, or at least argues for a fundamentally different set of constraints.

Prospect Theory and Framing Effects

Consider the following pair of choices (Tversky and Kahneman 1981): an unusual disease is expected to kill 600 people. Two alternative programmes to combat the disease have been proposed.

Programme A will save 200; with Programme B, there is a one-third probability that 600 people will be saved, and a two-thirds probability that no one will be saved.

Next, consider the following restatement of what would appear to be the same options:

If Programme C is adopted 400 people will die; with Programme D, there is a one-third probability that nobody will die, and a two-thirds probability that 600 people will die.

Among subjects given a choice between A and B, most choose the safe option A, while the majority of the subjects facing the second pair of choices choose the risky option D.

Kahneman and Tversky's (1979) prospect theory combines issues discussed in sections "The Independence Axiom" and "Evolution and Learning", with a more general critique of standard economic models, or at least of how such models are used in practice. Thus, while a standard model might favour a level of abstraction that ignores the framing issue above, Kahneman and Tversky (1979) argue that identifying the particular frame that the individual is likely to confront should be central to decision theory. In particular, these authors focus on the differential treatment of gains and losses. Prospect theory defines preferences not over lotteries of terminal wealth but over gains and losses, measured as differences from a status quo. In applications, the status quo is identified in a variety of ways.

For example, Köszegi and Rabin (2005–6) provide a theory of the status quo and utilize the

resulting model to study a monopoly problem. In their theory, the DM's optimal choice becomes the status quo. Thus, the simplest form of the Köszegi–Rabin model defines optimal choices from a set A as $C(A) = \{x \in A | U(x,x) \geq U(y,x) \; \forall \; y \in A\}$. Hence, $x \in A$ is deemed to be a possible choice from A if the DM who views x as his reference point does not strictly prefer some other alternative y.

The three lines of work discussed below all represent a fundamental departure from the standard modelling of economic decisions: they describe behaviour as the outcome of a game even in a single person problem.

Preference Reversals

Strotz (1955–6) introduces the idea of dynamic inconsistency: the possibility that a DM may prefer to consume x in period 2 to consuming y in period 1, if he makes the choice in period 0, but may have the opposite preference if he makes the choice in period 1. Strotz suggests that the appropriate way to model dynamically inconsistent behaviour is to assume that the period 0 individual treats his period 1 preference (and the implied behaviour) as a constraint on what he can achieve. Thus, suppose the period 0 DM has a choice between committing to z for period 2 consumption, or rejecting z and giving his period 1 self the choice between x in period 2 and y in period 1. Suppose also that the period 0 self prefers x to z and z to y while the period 1 self prefers y to x. Then, the Strotz model would imply that the DM ends up consuming z in period 2: the period 0 self realizes that if he does not commit to z, his period 1 self will choose y over x, which, for the period 0 self, is the least desirable outcome. Therefore, the period 0 self will commit to z. Hence, dynamic inconsistency leads to a preference for commitment.

Peleg and Yaari (1973) propose to reconcile the conflict among the different selves of a dynamically inconsistent DM with a strategic equilibrium concept. Their reformulation of Strotz's notion of consistent planning has facilitated the application of Strotz's ideas to more general settings, including dynamic games.

Imperfect Recall

An explicit statement of the perfect recall assumption and analysis of its consequences (Kuhn 1953) is one of the earliest contributions of extensive form game theory. In contrast, the analysis of forgetfulness, that is, extensive form games where the individual forgets his own past actions or information, is relatively recent (Piccione and Rubinstein 1997).

Piccione and Rubinstein observe that defining optimal behaviour for players with imperfect recall is problematic and propose a few alternative definitions (1997). Subsequent work has focused on what they call the multi-selves approach. In the multi-selves approach to imperfect recall, as in dynamic inconsistency, each information set is treated as a separate player. Optimal behaviour is a profile of behavioural strategies and beliefs at information sets such that the beliefs are consistent with the strategy profile and each behavioural strategy maximizes the corresponding agent's payoff given his beliefs and the behaviour of the remaining agents. Hence, the multi-selves approach leads to a prediction of behaviour that is analogous to perfect Bayesian equilibrium.

Psychological Games

Harsanyi (1967–8) introduces the notion of a type to facilitate analysis of the interaction of players' information in strategic problems. He argues that the notion of a type is flexible enough to accommodate all uncertainty and asymmetric information that is relevant in games. Geanakoplos et al. (1989) observe that if payoffs are 'intrinsically' dependent on beliefs and beliefs are determined in equilibrium, then types cannot be defined independently of the particular equilibrium outcome. Their notion of a psychological game and type (for psychological games) allows for this interdependence between equilibrium expectations and payoffs.

Gul and Pesendorfer (2006) offer an alternative framework for dealing with interdependent preferences. In their analysis, players care not only about the physical consequences of their actions on their opponents, but also about their opponents' attitudes towards such consequences, and their opponents' attitudes towards others'

attitudes towards such consequences, and so on. Gul and Pesendorfer provide a model of interdependent preference types similar to Harsanyi's interdependent belief types to analyse situations in which preference interdependence may arise not from the interaction of (subjective) information but from the interaction of the individuals' attitudes towards the well-being of others.

Neuroeconomics

The most comprehensive challenge to the standard economic modelling of the individual comes from research in neuroeconomics. Neuroeconomists argue that no matter how much the standard conventions are expanded to accommodate behavioural phenomena, it will not be enough: understanding economic behaviour requires studying the physiological, and in particular, neurological mechanisms behind choice. Recent experiments relate choice-theoretic variables to levels of brain activity, the type of choices to the parts of the brain that are engaged when making these choices, and hormone levels to behaviour (Camerer 2007) provide a concise summary of recent research in neuroeconomics).

Neuroeconomists contend that 'neuroscience findings raise questions about the usefulness of some of the most common constructs that economists commonly use, such as risk aversion, time preference, and altruism' (Camerer et al. 2005). They argue that neuroscience evidence can be used directly to falsify or validate specific hypotheses about behaviour. Moreover, they claim that organizing choice theory and game theory around the abstractions of neuroscience will lead to better theories. Thus, neureconomics proposes to change both the language of game theory and what constitutes its evidence.

Conclusion

The interaction of behavioural economics and game theory has had two significant effects: first, it has broadened the subject matter and set of acceptable approaches to strategic analysis. New modelling techniques such as equilibrium notions that explicitly address biases have become acceptable and new questions such as the effect of ambiguity aversion in auctions have gained interest. More importantly, behavioural approaches have altered the set of empirical benchmarks – the stylized facts – that game theorists must address as they interpret their own conclusions.

See Also

- ▶ Allais Paradox
- ▶ Altruism in Experiments
- ▶ Ambiguity and Ambiguity Aversion
- ▶ Learning and Evolution in Games: An Overview
- ▶ Prospect Theory
- ▶ Preference Reversals
- ▶ Neuroeconomics

Bibliography

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole Americaine. *Econometrica* 21: 503–546.

Aumann, R.J., and A. Brandenburger. 1995. Epistemic conditions for Nash equilibrium. *Econometrica* 63: 1161–1180.

Camerer, C.F., G. Loewenstein, and D. Prelec. 2005. Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature* 43: 9–64.

Camerer, C.F. 2007. Neuroeconomics: Using neuroscience to make economic predictions. *Economic Journal* 117: C26–C42.

Crawford, V.P. 1990. Equilibrium without independence. *Journal of Economic Theory* 50: 127–154.

Ellsberg, D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Eyster, E., and M. Rabin. 2005. Cursed equilibrium. *Econometrica* 73: 1623–1672.

Fehr, E., and K. Schimdt. 1999. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114: 817–868.

Geanakoplos, J., D. Pearce, and E. Stacchetti. 1989. Psychological games and sequential rationality. *Games and Economic Behavior* 1: 60–80.

Gul, F., and W. Pesendorfer. 2001. Temptation and self-control. *Econometrica* 2001: 1403–1435.

Gul, F., and W. Pesendorfer. 2006. *The canonical type space for interdependent preferences*, Working paper. Princeton: Princeton University.

Harsanyi, J. 1967–8. Games with incomplete information played by Bayesian players. *Management Science* 14: 159–182, 320–334, 486–502.

Ho, T., C. Camerer, and K. Weingelt. 1998. Iterated dominance and iterated best responses in p-beauty contests. *American Economic Review* 88: 947–969.
Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–292.
Köszegi, B., and M. Rabin. 2006. A model of reference-dependent preferences. *Quarterly Journal of Economics* 121: 1133–1166.
Kreps, D.M. 1979. A preference for flexibility. *Econometrica* 47: 565–576.
Kreps, D.M., and E.L. Porteus. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46: 185–200.
Kuhn, H.W. 1953. Extensive games and the problem of information. In *Contributions to the Theory of Games*, vol. 2. Princeton: Princeton University Press.
Machina, M. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27: 1622–1668.
O'Donoghue, T., and M. Rabin. 1999. Doing it now or later. *American Economic Review* 89: 103–124.
Peleg, B., and M.E. Yaari. 1973. On the existence of a consistent course of action when tastes are changing. *Review of Economic Studies* 40: 391–401.
Piccione, M., and A. Rubinstein. 1997. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior* 20: 3–24.
Strotz, R.H. 1955–6. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23: 165–180.
Tversky, A., and D. Kahnemann. 1974. Judgment under uncertainty: Heuristic and biases. *Science* 85: 1124–1131.
Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211: 453–458.

Behavioural Finance

Robert Bloomfield

Abstract

Behavioural finance began as an attempt to understand why financial markets react inefficiently to public information. One stream of behavioural finance examines how psychological forces induce traders and managers to make suboptimal decisions, and how these decisions affect market behaviour. Another stream examines how economic forces might keep rational traders from exploiting apparent opportunities for profit. Behavioural finance remains controversial, but will become more widely accepted if it can predict deviations from traditional financial models without relying on too many ad hoc assumptions, and expand to settings (particularly corporate finance) in which arbitrage forces are weaker.

Keywords

Accruals anomaly; Anomalies; Arbitrage; Behavioural finance; Book-to-market effect; Capital asset pricing model; Efficient markets hypothesis; Equity premium puzzle; Gambler's fallacy; Home bias puzzle; Hot-hand fallacy; Incomplete revelation hypothesis; Limited attention; Market microstructure; Momentum; Miscalibration; Mispricing; Overconfidence; Pattern recognition; Post-earnings-announcement drift; Prospect theory; Reversal; Risk; Size effect

JEL Classifications

C9

Mounting evidence suggests that a variety of trading strategies generate returns that are larger than permitted by the reigning theory of efficient financial markets. Defenders of efficient markets theory argue that the anomalies represent methodological errors, and in many cases they appear to have been correct. In cases where the anomalies appear robust, the debates turn to two other questions. First, why would investors make systematic trading errors that could result in mispricing? Second, why wouldn't smarter traders exploit those errors, thereby driving prices to appropriate levels? Many answers to the first question have relied heavily on the branch of psychology called 'behavioural decision theory', which has led to the entire body of research being dubbed 'behavioural finance' even though there is rarely much behavioural content in the literatures identifying pricing anomalies and explaining why price errors are not eliminated by smarter traders.

The next section of this article discusses the empirical evidence that market prices deviate from levels that would reflect perfectly rational

traders acting in competitive markets (the 'anomalies' literature). I then discuss literatures that document how behavioural forces can explain these anomalies, and that examine why irrational traders might influence prices in competitive markets. I conclude by suggesting some promising future directions in behavioural finance.

Anomalies

In 1968, two accounting professors reported that markets react sharply to earnings announcements over the course of a few days, and then continue drifting in the same direction for the better part of a year (Ball and Brown 1968). This post-earnings-announcement drift (PEAD) appeared to provide an easy opportunity for making money: one could create a hedged portfolio that is long in firms that have just announced good news and short in firms that have just announced bad news, so that it earns positive returns from no net investment.

The fact that prices react at all to earnings was surprising enough, given that earnings was then viewed as an accounting fiction describing past events, with no bearing on the future cash flows of the firm that should entirely determine firm value. (Accounting 'fictions' like earnings and book value are now known to provide important information about future cash flows, spawning a large field of financial accounting research.) But the subsequent drift was even more surprising, as it flew in the face of the recently developed efficient markets hypothesis (EMH), subsequently codified by Gene Fama (1970). The EMH relies on competition among investors to assert that strategies based on public information cannot earn returns after adjusting for risk. If all investors know that holding the PEAD portfolio would allow for excess returns, they would compete to hold the portfolio, and drive prices to the level needed to eliminate those returns.

PEAD has turned out to be one of the first – and most robust – of a large number of market anomalies. Initial explanations for PEAD were that the predictable returns simply reflect the expected returns that investors demand to compensate for the risk the PEAD portfolio would impose on them. Such arguments were made much more difficult by Bernard and Thomas (1990), who showed that about half the returns to the PEAD portfolio were experienced in the three-day windows surrounding the two subsequent earnings announcements. Thus, any risk-based explanation would require firms with extremely good or bad earnings news to experience dramatic changes in systematic risk for only a few days a year, several months in the future. The alternative explanation, proffered by Bernard and Thomas, was that investors simply did not understand the implications of current earnings for future earnings – an assertion that has been repeatedly supported by studies of analysts' earnings estimates and laboratory experiments. Researchers were successful enough in ruling out the risk explanation, and in tying future returns to the information content of current earnings, so that Fama (1998, p. 304) concluded that PEAD 'has survived robustness checks', and was possibly 'above suspicion'.

Three other robust anomalies seem more likely to reflect compensation for risk than mispricing: the book-to-market effect, the size effect and the momentum effect. The book-to-market ratio is the ratio of a firm's net assets (as reported on the firm's balance sheet) to the total market value of the firm's outstanding stock. Firms with low book-to-market ratios earn substantially higher returns than those with high book-to-market ratios (the book-to-market effect), as if the market value reverts over time to the value indicated by the accounting statements. Firms with small market capitalization earn higher returns than firms with large market capitalization (the size effect), as if small firms are consistently underpriced. Stocks that move strongly upwards or downwards over a three- to six-month period are very likely to continue moving in that direction over a subsequent three to six months (the momentum effect), as if the market responds slowly to changes in value.

Distinguishing risk and mispricing is difficult for book-to-market and size and momentum effects because researchers have no hypothesis that the mispricing will be corrected at some particular moment. (In contrast, the theory explaining PEAD suggests that mispricing will be revealed and corrected upon subsequent earnings announcements). Proponents of efficient markets have

provided evidence that book-to-market and size capture systematic risk, and have expanded the traditional asset pricing model to include book-to-market, size and (less frequently) momentum as risk factors. However, analysts appear to view book-to-market as an indicator of mispricing rather than risk, as indicated by examinations of analyst reports and controlled experiments.

Researchers in finance and accounting have identified a host of other pricing anomalies. Here is a selective sampling of some of the most well known, all of which remain controversial:

- *Long-term price reversal*. Stocks that move strongly over a three- to five-year period are very likely to reverse a portion of those movements over a following three- to five-year period (DeBondt and Thaler 1985). Evidence for long-term reversal tends to be more controversial than evidence for short-term momentum, because longer horizons make it harder to guarantee appropriate computation of risk-adjusted returns.
- *The equity premium puzzle*. A diversified portfolio of equity securities should earn higher returns than a portfolio of bonds, because of the additional risk equities impose on investors. However, the equity premium appears far too large relative to the associated risk (Mehra and Prescott 1985).
- *The home bias puzzle*. Both institutional and individual investors tend to hold a disproportionate amount of their portfolios in firms based in their own countries and regions. This may reflect a bias to purchase familiar stocks (Huberman 2001), or the inside information held by local investors (Coval and Moskowitz 2001).
- *Excessive volatility and excessive volume.* Shiller (1981) has argued that market prices are excessively volatile, relative to the volatility of fundamentals. Many others, including Kandel and Pearson (1995), have argued that trade volume is far too high to be explained by traditional theory, in light of the Milgrom and Stokey (1982) 'no-trade theorem', which proves that, in the absence of non-informational motivations for trade, such as a need for liquidity or sharing of risk, markets should not include any trade.
- *The accruals anomaly.* Firms' earnings can be decomposed into cash flows and accruals (defined as earnings minus cash flows). Sloan (1996) showed that firms with large positive accruals earn lower future returns than firms with large negative accruals, as if investors are unaware that accruals – which do not represent cash flows and are easily manipulated by managers – reverse rapidly.

Individual Behaviour

The variety of market anomalies has led some to doubt the validity of the EMH, but few researchers are likely to let go of the efficient markets perspective without a coherent and parsimonious theory of when to predict which types of anomalies. One branch of psychology, called 'behavioural decision theory' (BDT), appears particularly well-suited to imposing regular structure on otherwise ad hoc results. BDT researchers have shown that a variety of apparently irrational behaviours can be explained by a relatively parsimonious set of theories. For their part, behavioural finance researchers have sought to use empirical and experimental studies to show that behavioural theories can describe the actions of individual investors (as well as managers), and to use theoretical methods to show that a small set of behavioural theories can account for the wide variety of market anomalies. Four streams of results feature most prominently in behavioural finance: prospect theory, miscalibration, pattern recognition and limited attention.

Prospect Theory

Throughout the 1970s, Amos Tversky and Daniel Kahneman published a series of papers characterizing how people value outcomes. This research ultimately resulted in a mathematical representation of subjective (hedonic) value called 'prospect theory' (Kahneman and Tversky 1979), for which Kahneman won the 2002 Nobel Prize in

economics (Amos Tversky died in 1996). Prospect theory emphasizes three features of the value function: that the hedonic value of an outcome is determined by whether the outcome is a gain or loss relative to the agent's reference point; that the negative hedonic value of a loss more than offsets the positive hedonic value of a gain of the same size; and that the marginal effect of increasing a gain (or loss) is decreasing in the size of the gain (or loss).

Prospect theory yields a variety of predictions that describe individual behaviour well, and that can also account for several market anomalies. Prospect theory helps to explain a common behaviour termed the 'disposition effect' (Shefrin and Statman 1985) – traders will close out profitable investments quickly, to lock in gains, while holding on to their losing investments or perhaps even invest more in them, in hopes that the investment will turn around. Let us assume that a trader has bought a stock at 50 dollars, and that it is now priced at 80 dollars. Using the 50-dollar purchase price as a reference point, the trader has a 30-dollar gain, and (because the marginal effect of increasing a gain is decreasing in the size of the gain) the agent is risk-averse, and will want to close the position quickly to avoid risk. If the price fell to 20 dollars, however, the trader has a 30-dollar loss, and (because the marginal effect of increasing a loss is decreasing in the size of the loss) the agent is risk seeking, and will want to keep the position open to take on more risk.

Terry Odean (1998a) has shown clear evidence of the disposition effect among thousands of individual investors at a brokerage firm. Unfortunately for the investors, selling winners and holding on to losers is nearly the opposite of the profitable momentum strategy, which involves buying recent winners and selling recent losers. As a result, the stocks the investors held subsequently underperformed the stocks they sold. The disposition effect does not seem restricted to amateurs. Coval and Shumway (2005) show that professional commodity traders who have net losses near the end of the day tend to trade quite aggressively until trading closes, and take on significant risk. Finally, Frazzini (2006) ties the disposition effect back to price anomalies by providing evidence that disposition effects drive short-term momentum, because the relatively rapid selling of winners slows reactions to good news, while the tendency to hold losers slows reactions to bad news.

The disposition effect is driven by the different curvatures of the value function in the loss and gain realms. Curvature is important when investors evaluate the risk of relatively small changes in wealth. Investors who evaluate the risk of large wealth changes are influenced instead by the different average slopes of the value function in the loss and gain realms. Because the average slope is flatter in the realm of gains, investors with large gains in hand are likely to appear less risk-averse than those with losses or small gains. Evidence from experiments (Thaler and Johnson 1990) and game show contestants (Gertner 1993) are consistent with this 'house money' effect, named after the exaggerated risk tolerance of the behaviour of gamblers who have won money from the house, and therefore are risking only the house's money. Barberis et al. (2001) show that the house money effect can account for both short-term momentum and long-term reversal. Short-term momentum arises because traders demand more compensation for risk after price declines, further depressing prices, while demanding less compensation for risk after price increases, further inflating prices. Similar reasoning shows that the house money effect can account for the book-to-market effect and an exaggerated equity premium.

While prospect theory is a relatively parsimonious and powerful theory, its predictions are highly sensitive to assumptions about how people identify benchmarks against which to measure gains and losses, and under what circumstances they might evaluate gains and losses of portfolios, rather than of individual securities. The field of 'mental accounting' (Barberis et al. 2006) addresses such questions.

Miscalibrated Confidence

Financial models of trade traditionally assume that agents have confidence calibrated to reflect the precision of their information. Experiments

show that people rarely satisfy this requirement. People tend to be overconfident in their ability to predict events when they have very poor information, while people who are asked easy questions tend to be underconfident. Psychologists call this tendency the 'hard–easy' effect (Griffin and Tversky 1992); Bloomfield et al. (2000) call it 'moderated confidence' because confidence is moderated from the optimal level towards a prior belief of moderate data reliability, as if people are rational Bayesians with imperfect information about the reliability of their data.

Because financial outcomes are so hard to predict, people are likely to be overconfident, rather than under-confident. Indeed, evidence of overconfidence is widespread. Odean (1999) finds that individual investors trade far too frequently, apparently overconfident in their ability to identify mispriced securities. Malmendier and Tate (2005) find that many executives are overconfident in their firms' futures (as evidenced by their failure to exercise stock options before expiration), and further show that more overconfident executives are more likely to engage in value-reducing mergers.

Theoretical and experimental research has shown that calibration errors can account for a variety of known anomalies. Gervais and Odean (2001) and Odean (1998b) examine how overconfidence can lead to excessive trading. Daniel et al. (1998) show that overconfidence can account for both overreactions and underreactions to information. In a similar vein, Bloomfield et al. (2003) show that overconfident inferences from old earnings numbers, which have little information content once newer numbers are available, lead to both post-earnings-announcement drift and overreactions to earnings trends.

Pattern Recognition

The human mind has a gift for finding order in chaos, even when objective analysis shows no order to be found. In such cases, people show remarkable consistency in the order they perceive. People fall prey to the gambler's fallacy when they expect that a coin that has come up 'heads' many times in a row is then more likely to come up 'tails' because such streaks are typically short-lived. People fall prey to the 'hot-hand' fallacy when they mistakenly believe that basketball players who have made ten free throws in a row are especially likely to make the next, even though this is not the case (a professional basketball player's free throw performance is not distinguishable from a random series with a constant mean). The tendency to see patterns in random sequences is likely to be particularly important in financial markets, where competitive pressures force market prices to follow a random walk (after risk premia are accounted for). Despite the randomness in stock movements, many investors subscribe to 'technical analysis' trading strategies (and expensive newsletters) based on elaborate patterns like 'head and shoulders' and 'cup with handle', even though systematic research has found little evidence that such patterns can predict future stock movements.

Barberis et al. (1998) claim that people who observe a random walk are likely to fluctuate between beliefs in the gambler's fallacy (in which any trends are quickly reversed) and beliefs in the hot hand (in which trends continue), depending on how many reversals in price they have seen in recent periods. They then prove that such beliefs can account for both short-term price momentum and long-term price reversal. Bloomfield and Hales (2002) find experimental support for that assumption.

Limited Attention

A fundamental tenet of cognitive science is that people have limited cognitive resources, implying that their attention to financial information and investment opportunities may be determined by economically irrelevant factors such as how information is presented or how often it is talked about by others. Experiments have found that even experienced analysts draw conclusions that are coloured by seemingly irrelevant aspects of how financial information is presented (Hirst and Hopkins 1998). Employees' decisions on how to invest their defined contribution pension funds are

dramatically influenced by how the options are presented (Benartzi and Thaler 2001), while their decision to enrol in such plans at all are dramatically increased by a policy that makes investment the default option, so that enrolment requires no attention at all (Benartzi and Thaler 2004).

Limited attention may determine how stocks come in and out of favour, and provides a natural explanation for the home bias puzzle – people naturally notice local firms more readily than distant firms. Limited attention may also explain the tendency of firms to attract attention (and trading volume) when their earnings are growing rapidly, but be ignored when they perform poorly for long periods. Lee and Swaminathan (2000) argue that such tendencies might explain short-term momentum, and support their argument by showing that firms with low volume and strong returns show strong momentum in returns (as if they are underpriced while still neglected), while those with high volume and strong returns show long-term reversal (as if they are overpriced at the peak of attention).

Accounting researchers have been particularly interested in the effects of limited attention, because they may explain why people care so much about accounting regulations that alter only how information is presented, and not the information content of the complete accounting disclosure. A highly publicized example is the controversy over whether employee stock option costs should be deducted from reported earnings per share; in both cases, investors could gather all relevant information from the footnotes to the financial statements. Bloomfield (2002) argues that fewer investors attend to footnotes than to earnings, and that standard models of information aggregation predict that market prices less completely reveal information that is held by fewer investors – a result repeatedly confirmed in laboratory markets. This 'incomplete revelation hypothesis' runs counter to the EMH, which is typically applied to all public information regardless of how it is presented. However, accounting researchers have made considerable progress in understanding how different presentation options, such as the formating, isolation and ordering of text can alter investors' attention to and weighting of the information in that text (see, for example, Maines and McDaniel 2000).

Limits to Arbitrage

Studies of individual behaviour show that investors and managers make systematic errors of judgement, but do not explain how other investors fail to exploit, and thereby eliminate, any aggregate mispricing.

A number of studies have noted that arbitrage may be limited by risks that cannot be captured as risk factors in traditional asset pricing models. Even if a pricing error must eventually converge (as when two securities representing claims on the same underlying assets have different prices), such convergence may not be rapid, and may even be preceded by additional divergence. While asset pricing models like the capital asset pricing model (CAPM) conclude that such idiosyncratic risk does not affect price levels, Pontiff (2006) has argued forcefully that idiosyncratic risk still hinders the correction of price errors by effectively imposing a 'holding cost' on arbitrageurs. Idiosyncratic risk restricts arbitrage most severely when a trader uses borrowed capital to engage in arbitrage, because a short-term loss may result in a margin call, or may lead the investors to infer that the arbitrageur has a poor strategy, and therefore withdraw their funds (Shleifer and Vishny 1997). DeLong et al. (1990) take these arguments one step further: they assume that the noise in returns is driven by irrational traders, and then show that these traders still earn sufficient returns for them to survive indefinitely.

Another line of literature notes that rational arbitrageurs might earn greater profits by exacerbating price errors rather than disciplining them. Abreu and Brunnermeier (2002) construct a model in which irrational traders drive prices too high, a fact that eventually becomes known to every arbitrageur. Because arbitrageurs do not know whether other arbitrageurs have yet learned of the overpricing, each one continues to 'ride the bubble' after they learn of the overpricing, rather than pop it, because they expect others to do so as well. As a result, the arbitrageurs continue

magnifying the bubble even after each individual arbitrageur knows that prices are too high.

The preceding explanations of limited arbitrage are largely devoid of behavioural content – the price errors that fail to be corrected could arise from any cause, including completely random trading. However, researchers do occasionally examine how specific biases can limit arbitrage opportunities. Overconfidence, in particular, has been shown to be difficult to arbitrage. For example, Kyle and Wang (1997) show that overconfident traders can effectively gain 'elbow room' in a market, just as a trader in a Cournot oligopoly game can benefit by committing to aggressive production, and forcing others to produce less. As a result, overconfident traders earn enough trading gains to persist.

Conclusion and Future Directions

This history of behavioural finance fits well within Kuhn's (1962) narrative of scientific revolution. Early researchers uncovered results that were anomalous within the paradigm of efficient markets; as they became convinced that the anomalies were not simply the result of methodological error, researchers sought a new paradigm that could encompass the anomalies, as well as the predictions of the traditional theory. This new paradigm assumes that markets include some participants who optimize their expected utility, along with others whose susceptibility to psychological forces leads them to behave suboptimally.

No behavioural alternative will ever rival the coherence, parsimony and power of traditional efficient markets theory, because psychological forces are too complex. Thus, behavioural researchers in finance must devote themselves to the 'normal science' suggested by their new paradigm: documenting and refining our understanding of how psychological forces influence individual behaviour in financial settings, and how those behaviours affect market phenomena. This will require much more attention to behavioural psychology than is evident in the existing body of research. (As of 2007, few papers in behavioural finance rely on psychological research published after the 1970s.) Perhaps more importantly, advances in behavioural finance will require more attention to the details of market microstructure, which influence individual behaviour, and how those behaviours affect market-level phenomena. Finally, researchers in behavioural finance can expand their scope beyond describing the behaviour of investors and prices in highly competitive asset markets. Behavioural theories are likely to have greater ability to explain phenomena in settings that provide fewer opportunities for others to exploit (and thereby eliminate) suboptimal outcomes. For example, decisions on how to hire and compensate executives, and on when and how to raise and invest capital, seem particularly susceptible to behavioural analysis (as in Shefrin 2005).

See Also

- ▶ Arbitrage
- ▶ Behavioural Economics and Game Theory
- ▶ Bubbles
- ▶ Efficient Markets Hypothesis
- ▶ Prospect Theory

Bibliography

Abreu, D., and M.K. Brunnermeier. 2002. Synchronization risk and delayed arbitrage. *Journal of Financial Economics* 66: 341–360.

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6: 159–178.

Barberis, N., A. Shleifer, and R. Vishny. 1998. A model of investor sentiment. *Journal of Financial Economics* 49: 307–343.

Barberis, N., M. Huang, and T. Santos. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116: 1–53.

Barberis, N., M. Huang, and R.H. Thaler. 2006. Individual preferences, monetary gambles and stock market participation: A case for narrow framing. *American Economic Review* 96: 1069–1090.

Benartzi, S., and R.H. Thaler. 2001. Naïve diversification strategies in defined contribution savings plans. *American Economic Review* 91: 79–98.

Benartzi, S., and R.H. Thaler. 2004. Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy* 112: S164–S187.

Bernard, V.L., and J. Thomas. 1990. Evidence that stock prices do not fully reflect the implications of current

earnings for future earnings. *Journal of Accounting and Economics* 13: 305–341.
Bloomfield, R. 2002. The incomplete revelation hypothesis: Implications for financial reporting. *Accounting Horizons* 16: 233–244.
Bloomfield, R., and J. Hales. 2002. Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs. *Journal of Financial Economics* 65: 397–415.
Bloomfield, R., R. Libby, and M.W. Nelson. 2000. Underreactions, overreactions and moderated confidence. *Journal of Financial Markets* 3: 113–137.
Bloomfield, R., R. Libby, and M.W. Nelson. 2003. Over-reliance on previous years' earnings. *Contemporary Accounting Research* 20: 1–31.
Coval, J., and T. Moskowitz. 2001. The geography of investment: Informed trading and asset prices. *Journal of Political Economy* 109: 811–841.
Coval, J.D., and T. Shumway. 2005. Do behavioral biases affect prices? *Journal of Finance* 60: 1–34.
Daniel, K., D. Hirshleifer, and A. Subrahmanyam. 1998. Investor psychology and security market under- and overreaction. *Journal of Finance* 53: 1839–1886.
DeBondt, W.F.M., and R.H. Thaler. 1985. Does the stock market overreact? *Journal of Finance* 40: 793–807.
DeLong, J.B., A. Shleifer, L.H. Summers, and R.J. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98: 703–738.
Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.
Fama, E.F. 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49: 283–306.
Frazzini, A. 2006. The disposition effect and underreaction to news. *Journal of Finance* 61: 2017–2046.
Gertner, R. 1993. Game shows and economic behavior: Risk taking on 'card sharks'. *Quarterly Journal of Economics* 151: 507–521.
Gervais, S., and T. Odean. 2001. Learning to be overconfident. *Review of Financial Studies* 14: 1–27.
Griffin, D., and A. Tversky. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology* 24: 411–435.
Hirst, D.E., and P.E. Hopkins. 1998. Comprehensive income reporting and analysts' valuation judgments. *Journal of Accounting Research* 36(Suppl): 47–75.
Huberman, G. 2001. Familiarity breeds investment. *Review of Financial Studies* 14: 659–680.
Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–292.
Kandel, E., and N.D. Pearson. 1995. Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy* 103: 831–872.
Kuhn, T.S. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Kyle, A., and F.A. Wang. 1997. Speculation duopoly with agreement to disagree: Can overconfidence survive the market test? *Journal of Finance* 52: 2073–2090.
Lee, C.M.C., and B. Swaminathan. 2000. Price momentum and trading volume. *Journal of Finance* 55: 2017–2033.
Maines, L.A., and L.S. McDaniel. 2000. Effects of comprehensive-income characteristics on non-professional investors' judgments: The role of financial- statement presentation format. *Accounting Review* 75: 179–207.
Malmendier, U., and G. Tate. 2005. CEO overconfidence and corporate investment. *Journal of Finance* 60: 2661.
Mehra, R., and E.C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15: 145–161.
Milgrom, P., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26: 17–27.
Odean, T. 1998a. Are investors reluctant to realize their losses? *Journal of Finance* 53: 1775–1798.
Odean, T. 1998b. Volume, volatility, price, and profit when all traders are above average. *Journal of Finance* 53: 1887–1934.
Odean, T. 1999. Do investors trade too much? *American Economic Review* 89: 1279–1298.
Pontiff, J. 2006. Costly arbitrage and the myth of idiosyncratic risk. *Journal of Accounting and Economics* 42: 35–52.
Shefrin, H. 2005. *Behavioral Corporate Finance*. New York: McGraw-Hill/Irwin.
Shefrin, H., and M. Statman. 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* 40: 777–790.
Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.
Shleifer, A., and R.W. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 52: 35–55.
Sloan, R. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review* 71: 289–315.
Thaler, R.H., and E.J. Johnson. 1990. Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science* 36: 643–660.

Behavioural Game Theory

Colin F. Camerer

Abstract

Behavioural game theory uses experimental regularities and psychology to model formally how limits on strategic thinking, learning, and social preferences interact when people actually play games. Emerging theories of

behaviour in ultimatum and trust games (and others) focus on an aversion to inequality, reciprocity, or concern for social image. Learning models often focus on numerical updating of an unobserved propensity to choose a strategy (including fictitious play updating of beliefs as a special case). Models of limits on strategic thinking assume players are in equilibrium, but respond with error, or there is a cognitive hierarchy of increasingly sophisticated reasoning.

Keywords

Adaptive expectations; Altruism; Analytical game theory; Auctions; Behavioural game theory; Bounded rationality; Cognitive hierarchy theories; Communication; Competition; Contracting; Coordination; Dictator games; Direction learning; Disequilibrium behaviour; Entry deterrence games; Experience weighted attraction; Individual learning in games; Inequality aversion; Learning; Mixed strategy equilibrium; Moral hazard; Negative reciprocity; Quantal response equilibrium; Rational expectations; Rawls, J.; Reciprocity; Replicator dynamics; Self-interest; Signalling; Social preferences; Tit for tat; Trust games; Ultimatum games; Utilitarianism; Weighted fictitious play

JEL Classifications

C7

Analytical game theory assumes that players choose strategies which maximize the utility of game outcomes, based on their beliefs about what others players will do, given the economic structure of the game and history; in equilibrium, these beliefs are correct. Analytical game theory is enormously powerful, but it has two shortcomings as a complete model of behaviour by people (and other possible players, including non-human animals and organizations).

First, in complex naturally occurring games, equilibration of beliefs is unlikely to occur instantaneously. Models of choice under bounded rationality, predicting initial choices and equilibration with experience, are therefore useful.

Second, in empirical work, only received (or anticipated) payoffs are easily measured (for example, prices and valuations in auctions, or currency paid in an experiment). Since games are played over *utilities* for received payoffs, it is therefore necessary to have a theory of social preferences – that is, how measured payoffs determine players' utility evaluations – in order to make predictions.

The importance of understanding bounded rationality, equilibration and social preferences is provided by hundreds of experiments showing conditions under which predictions of analytical game theory are sometimes approximately satisfied, and sometimes badly rejected (Camerer 2003). This article describes an emerging approach called 'behavioural game theory', which generalizes analytical game theory to explain experimentally observed violations. Behavioural game theory incorporates bounds on rationality, equilibrating forces, and theories of social preference, while retaining the mathematical formalism and generality across different games that has made analytical game theory so useful. While behavioural game theory is influenced by laboratory regularities, it is ultimately aimed at a broad range of applied questions such as worker reactions to employment terms, evolution of market institutions, design of auctions and contracts, animal behaviour, and differences in game-playing skill.

Social Preferences

Let us start with a discussion of how preferences over outcomes of game can depart from pure material self-interest. In an ultimatum game a Proposer is endowed with a known sum, say ten dollars, and offers a share to another player, the Responder. If the Responder rejects the offer they both get nothing. The ultimatum game is a building block of more complex natural bargaining and a simple tool to measure numerically the price that Responders will pay to punish self-servingly unfair treatment.

Empirically, a large fraction of subjects rejects low offers of 20 per cent or so. Proposers fear these rejections reasonably accurately, and make offers around 40 per cent rather than very small

offers predicted by perceived self-interest. (The earliest approximations of whether Proposers offer expected profit-maximizing offers, by Roth et al. 1991, suggested they did. However, those estimates were limited by the method of presenting Responders only with specific offers; since low offers are rare, it is hard to estimate the rejection rate of low offers accurately and hence hard to know conclusively whether offers are profit-maximizing. Different methods, and cross-population data used in Henrich et al. 2005, established that offers are too generous, even controlling for risk-aversion of the Proposers.) This basic pattern scales up to much higher stakes (the equivalent of months of wages) and does not change much when the experiment is repeated, so it is implausible to argue that subjects who reject offers (often highly intelligent college students) are confused.

It is crucial to note that rejecting two dollars out of ten dollars is a rejection of the *joint* hypothesis of utility-maximization and the auxiliary hypothesis that player *i's* utility depends on only her own payoff x_i. An obvious place to repair the theory is to create a parsimonious theory of social preferences over (x_i, x_j) (and possibly of other features of the game) which predicts violations of self-interest across games with different structures. I will next mention some other empirical regularities, then turn to a discussion of such models of these regularities.

In ultimatum games, it appears that norms and judgements of fairness can depend on context and culture. For example, when Proposers earn the right to make the offer (rather than respond to an offer) by winning at a pre-play trivia game, they feel entitled to offer less – and Responders seem to accept less (Hoffman et al. 1994). Two comparative studies of small-scale societies show interesting variation across cultures. Subjects in a small Peruvian agricultural group, the Machiguenga, offer much less than those in other cultures (typically 15–25 per cent) and accept low offers. Across 15 societies, equality of average offers is positively related to the degree of cooperation in economic activity (for example, do men hunt collectively?) and to the degree of impersonal market trading (Henrich et al. 2005).

Ultimatum games tap negative reciprocity or vengeance. Other games suggest different psychological motives which correspond to different aspects of social preferences. In dictator games, a Proposer simply dictates an allocation of money and the Responder must accept it. In these games, Proposers offer less than in ultimatum games (about 15 per cent of the stakes on average), but offers vary widely with contextual labels and other variables (Camerer 2003, ch. 2). In trust games, an Investor risks some of her endowment of money, which is increased by the experimenter (representing a return on social investment) and given to an anonymous Trustee. The Trustee pays back as much of the increased sum as she likes to the Investor (perhaps nothing) and keeps the rest. Trust games are models of opportunities to gain from investment with no legal protection against moral hazard by a business partner. Self-interested Trustees will never pay back money; self-interested Investors with equilibrium beliefs will anticipate this and invest nothing. In fact, Investors typically risk about half their money, and Trustees pay back slightly less than was risked (Camerer 2003, ch. 2). Investments reflect expectations of repayment, along with altruism toward Investors (Ashraf et al. 2006) and an aversion to 'betrayal' (Bohnet and Zeckhauser 2004). Trustee payback is consistent with positive reciprocity, or a moral obligation to repay a player who risked money to benefit the group.

Importantly, competition has a strong effect in these games. If two or more Proposers make offers in an ultimatum game, and a single Responder accepts the highest offer, then the only equilibrium is for the Proposers to offer almost all the money to the Responder (the *opposite* of the prediction with one Proposer). In the laboratory this Proposer competition occurs rapidly, resulting in a very unfair allocation – almost no earnings for Proposers (for example, Camerer and Fehr 2006). Similarly, when there is competition among Responders, at least one Responder accepts low offers and Proposers seem to anticipate this effect and offer much less. These regularities help explain an apparent paradox, why the competitive model based on self-interest works so well in explaining market prices in experiments

with three or more traders on each side of the market. In these markets, traders with social preferences cannot make choices which reveal a trade-off of self-interest and concern for fairness. The parsimonious theory in which agents have social preferences can therefore explain both fairness-type effects in bilateral exchange and the absence of those effects in multilateral market exchange.

A good social preference theory should explain all these facts: rejections of substantial offers in ultimatum games, lower Proposer offers in dictator games than in ultimatum games, trust and repayment in trust games, and the effects of competition (which bring offers closer to the equilibrium self-interest prediction).

In 'inequality-aversion' theories of social preference, players prefer more money and also prefer that allocations be more equal (judged by differences in payoffs – Fehr and Schmidt 1999 – or by deviations from payoff shares and equal shares – Bolton and Ockenfels 2000). In a related 'Rawlsitarian' approach, players care about a combination of their own payoffs, the minimum payoff (à la Rawls) and the total payoff (utilitarian) (Charness and Rabin 2002). These simple theories account relatively well for the regularities mentioned above across games, with suitable parameter values.

Missing from the inequality aversion and Rawlsitarian theories is a reaction to the intentions of players. Intentions seem to be important because players are much less likely to reject unequal offers that are created by a random device or third party than equivalently unequal offers proposed by a player who benefits from inequality (for example, Blount 1995; Falk et al. 2007). In reciprocity theories which incorporate intentions, player A forms a judgement about whether another player B has sacrificed to benefit (or harm) her (for example, Rabin 1993). A likes to reciprocate, repaying kindness with kindness, and meanness with vengeance. This idea can also explain the results mentioned above, and the effects of intentions shown in other studies.

A newer class of theories focused on 'social image' – that is, player A cares about whether another player B believes A adheres to a norm of fairness. For example, Dufwenberg and Gneezy (2000) show that Trustee repayments in a trust game are correlated with the Trustee's perception of what he or she thought the Investor expected to be repaid. These models hinge on delicate details of iterated beliefs (A's belief about B's belief about A's fairness), so they are more technically complicated but can also explain a wider range of results (see Bénabou and Tirole 2006; Dillenberger and Sadowski 2006). Models of this sort are also better equipped to explain deliberate avoidance of information. For example, in dictator games where the dictator can either keep nine dollars or can play a ten-dollar dictator game (knowing the Recipient will *not know* which path was chosen), players often choose the easy nine dollar payment (Dana et al. 2006). Since they could just play the ten-dollar game and keep all ten dollars, the ten-dollars sacrifice is presumably the price paid to avoid knowing that another person knows you have been selfish (see also Dana et al. 2007).

Social preference utility theories and social image concerns like these could be applied to explain charitable contribution, legal conflict and settlement, wage-setting and wage dispersion within firms, strikes, divorces, wars, tax policy, and bequests by parents to siblings. Explaining these phenomena with a single parsimonious theory would be very useful and important for policy and welfare economics.

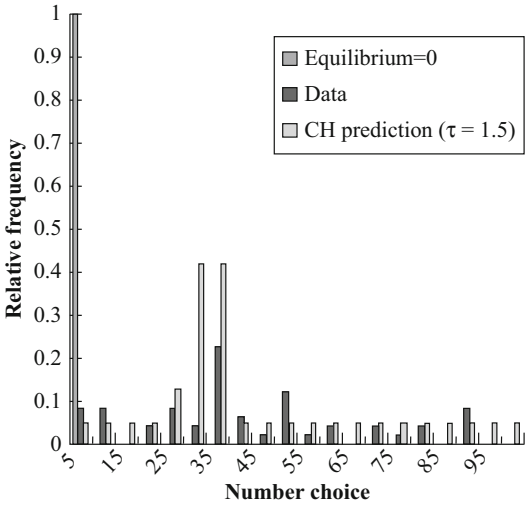
Limited Strategic Thinking and Quantal Response Equilibrium

In complex games, equilibrium analysis may predict poorly what players do in unique games, or in the first period of a repeated game. Disequilibrium behaviour is important to understand if equilibration takes a long time, and if initial behaviour is important in determining which of several multiple equilibria will emerge. Two types of theories are prominent: cognitive hierarchy theories of different limits on strategic thinking; and theories which retain the assumption of equilibrium beliefs but assume players make mistakes, choosing strategies with higher expected payoff deviations less often.

Cognitive hierarchy theories describe a 'hierarchy' of strategic thinking and constrain how the hierarchy works to make precise predictions. Iterated reasoning surely is limited in the human mind because of evolutionary inertia in promoting high-level thinking, because of constraints on working memory, and because of adaptive motives for overconfidence in judging relative skill (stopping after some steps of reasoning, believing others have reasoned less). Empirical evidence from many experiments with highly skilled subjects suggests that 0–2 steps of iterated reasoning are most likely in the first period of play. A simple illustration is the '*p*- beauty contest' game (Nagel 1995; Ho et al. 1998). In this game, several players choose a number in the interval [0,100]. The average of the numbers is computed, and multiplied by a value p (say 2/3). The player whose number is closest to p times the average wins a fixed prize.

In equilibrium players are never surprised what other players do. In the p-beauty contest game, this equilibrium condition implies that all players must be picking p times what others are choosing. This equilibrium condition only holds if everyone chooses 0 (the Nash equilibrium, consistent with iterated dominance). Figure 1 shows data from a game with $p = 7$ and compares the Nash prediction (choosing 0) and the fit of a cognitive hierarchy model (Camerer et al. 2004). In this game, some players choose numbers scattered from 0 to 100, many others choose p times 50 (the average if others are expected to choose randomly) and others choose p^2 times 50. When the game is played repeatedly with the same players (who learn the average after each trial), numbers converge toward zero, a reminder that equilibrium concepts do reliably predict where an adaptive process leads, even if they do not predict the starting point of that process.

In cognitive hierarchy theories, players who do k steps of thinking anticipate that others do fewer steps. Fully specifying these theories requires specifying what 0-step players do, what higher-step players think, and the statistical distribution of players' thinking levels. One type of theory assumes players who do k steps of thinking believe others do k-steps (Nagel 1995; Stahl and Wilson 1995; Costa-Gomes et al. 2001). This specification is analytically tractable (especially in games with $n >$ two players) but implies that as players do more thinking their beliefs are further from reality. Another specification assumes increasingly rational expectations – k-level players truncate the actual distribution $f(k)$ of k-step thinkers and guess accurately the relative proportions of thinkers doing 0 to $k - 1$ steps of thinking. Camerer et al. (2004) and earlier studies show how these cognitive hierarchy theories can fit experimental data from a wide variety of games, with similar thinking-step parameters across games.



Behavioural Game Theory, Fig. 1 Number choices and theoretical predictions in beauty contest games. *Note*: Players choose numbers from 0 to 100 and the closest number to 0.7 times the average wins a fixed prize. *Source*: Camerer and Fehr (2006)

These cognitive hierarchy theories ignore the benefits and costs of thinking hard. Costs and benefits can be included by relaxing Nash equilibrium, so that players respond stochastically to expected payoffs and choose better responses more often then worse ones, but do not maximize. Denote player i's beliefs about the chance that other players j will choose strategy k by $P_i(s_j^k)$. The expected payoff of player i's strategy s_i^h is $E(s_i^h) = \sum_k P_i\left(s_j^k\right)\pi_i\left(s_j^h, s_k^h\right)$ (where $\pi_i(x,y)$ is i's payoff if i plays x and j plays y). If player

i responds with a logit choice function, then $P_i\left(s_j^h\right) = exp\left(\lambda E\left(s_i^h\right)\right)/\sum_k exp\left(\lambda E\left(s_i^k\right)\right)$. In this kind of 'quantal response' equilibrium (QRE), each player's beliefs about choice probabilities of others are consistent with actual choice probabilities, but *players do not always choose the highest expected payoff strategy (and λ parameterizes the degree of responsiveness; larger λ implies better response)*. QRE fits a wide variety of data better than Nash predictions (McKelvey and Palfrey 1995, 1998; Goeree and Holt 2001). It also circumvents some technical limits of Nash equilibrium because players always tremble but the degree of trembling in strategies is linked to expected payoff differences.

Learning

In complex games, it is unlikely that equilibrium beliefs arise from introspection or communication. Therefore, theorists have explored the mathematical properties of various rules under which equilibration might occur when rationality is bounded.

Much research is focused on population evolutionary rules, such as replicator dynamics, in which strategies which have a payoff advantage spread through the population (for example, Weibull 1995). Schlag and Pollock (1999) show a link between imitation of successful players and replicator dynamics.

Several individual learning rules have been fit to many experimental data-sets (see ▶ Individual Learning in Games). Most of these rules can be expressed as difference equations of underlying numerical propensities or attractions of stage-game strategies which are updated in response to experience. The simplest rule is choice reinforcement, which updates chosen strategies according to received payoffs (perhaps scaled by an aspiration level or reference point). These rules fit surprisingly well in some classes of games (for example, with mixed strategy equilibrium, so that all strategies are played and reinforced relatively often) and in environments with little information, where agents must learn payoffs from experience, but can fit quite poorly in other games. A more complex rule is weighted fictitious play (WFP), in which players form beliefs about what others will do in the future by taking a weighted average of past play, and then choose strategies with high expected payoffs given those beliefs (Cheung and Friedman 1997). Camerer and Ho (1999) showed that WFP with geometrically declining weights is mathematically equivalent to generalized reinforcement in which unchosen strategies are reinforced as strongly as chosen ones. Building on this insight, they create a hybrid called experience weighted attraction (EWA). The original version of EWA has many parameters because it includes all the parameters used in the various special cases it hybridizes. The EWA form fits modestly better in some games (it adjusts carefully for overfitting by estimating parameters on part of the data and then forecasting out-of-sample), especially those with rapid learning across many strategies (such as pricing). In response to criticism about the number of free parameters, Ho et al. (2007) created a version with zero *learning parameters (just a response sensitivity λ as in QRE)* by replacing parameters by 'self-tuning' functions of experience.

Some interesting learning rules do not fit neatly into the class of strategy- updating difference equations. Often it is plausible to think that players are reinforcing learning *rules* rather than strategies (for example, updating the reinforcement rule or the WFP rule; see Stahl 2000). In many game it is also plausible that people update history-dependent strategies (like tit for tat; see Erev and Roth 2001; McKelvey and Palfrey 2001). Selten and Buchta (1999) discuss a concept of 'direction learning' in which players adjust based on experience in a 'direction' when strategies are numerically ordered.

All the rules described above are naive (called 'adaptive') in the sense that they do not incorporate the fact that other players are learning. Models which allow players to be 'sophisticated' and anticipate learning by other players (Stahl 1999; Chong et al. 2006) often fit better, especially with experienced subjects. Sophistication is particularly important if players are matched

together repeatedly – as workers in firms, firms in strategic alliances, neighbours, spouses, and so forth. Then players have an incentive to take actions that 'strategically teach' an adaptive player what to do. Models of this sort have more moving parts but can explain some basic stylized facts (for example, differences in repeated-game play with fixed 'partner' and random 'stranger' matching of players) and fit a little better than equilibrium reputational models in trust and entry deterrence games (Chong et al. 2006).

Conclusion

Behavioural game theory uses intuitions and experimental evidence to propose psychologically realistic models of strategic behaviour under rationality bounds and learning, and incorporates social motivations in valuation of outcomes. There are now many mathematical tools available in both of these domains that have been suggested by or fit closely to many different experimental games: cognitive hierarchy, quantal-response equilibrium, many types of learning models (for example, reinforcement, belief learning, EWA and self-tuning EWA), and many different theories of social preference based on inequality aversion, reciprocity, and social image. The primary challenge in the years ahead is to continue to compare and refine these models – in most areas, there is still lively debate about which simplifications are worth making, and why – and then apply them to the sorts of problems in contracting, auctions, and signalling that equilibrium analysis has been so powerfully applied to.

A relatively new challenge is to understand communication. Hardly any games in the world are played without some kind of pre-play messages (even in animal behaviour). However, communication is so rich that understanding how communication works by pure deduction is unlikely to succeed without help from careful empirical observation. A good illustration is Brandts and Cooper (2007), who show the nuanced ways in which communication and incentives, together, can influence coordination in a simple organizational team game.

See Also

- ▶ Adaptive Expectations
- ▶ Experimental Economics
- ▶ Individual Learning in Games

Bibliography

Ashraf, N., I. Bohnet, and N. Piankov. 2006. Decomposing trust and trustworthiness. *Experimental Economics* 9: 193–208.

Bénabou, R., and J. Tirole. 2006. Incentives and prosocial behavior. *American Economic Review* 96: 1652–1678.

Blount, S. 1995. When social outcomes aren't fair – The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes* 63: 131–144.

Bohnet, I., and R. Zeckhauser. 2004. Trust, risk and betrayal. *Journal of Economic Behavior & Organization* 55: 467–484.

Bolton, G.E., and A. Ockenfels. 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review* 90: 166–193.

Brandts, J., and A. Cooper. 2007. It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association* 5(6): 1223–1268.

Camerer, C.F. 2003. *Behavioral game theory: Experiments on strategic interaction*. Princeton: Princeton University Press.

Camerer, C.F., and E. Fehr. 2006. When does 'economic man' dominate social behavior? *Science* 311: 47–52.

Camerer, C.F., and T.H. Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* 67: 827–874.

Camerer, C.F., T.-H. Ho, and J.-K. Chong. 2004. A cognitive hierarchy model of games. *Quarterly Journal of Economics* 119: 861–898.

Charness, G., and M. Rabin. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117: 817–869.

Cheung, Y.-W., and D. Friedman. 1997. Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior* 19: 46–76.

Chong, J.-K., C.F. Camerer, and T.H. Ho. 2006. A learning-based model of repeated games with incomplete information. *Games and Economic Behavior* 55: 340–371.

Costa-Gomes, M., V.P. Crawford, and B. Broseta. 2001. Cognition and behavior in normal-form games: An experimental study. *Econometrica* 69: 1193–1235.

Dana, J., D.M. Cain, and R.M. Dawes. 2006. What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes* 100: 193–201.

Dana, J.D., R.A. Weber, and J.X. Kuang. 2007. Exploiting moral wiggle room: Behavior inconsistent with a preference for fair outcomes. *Economic Theory* 33(1): 67–80.
Dillenberger, D., and P. Sadowski. 2006. *Ashamed to be selfish*. Princeton: Princeton University Press.
Dufwenberg, M., and U. Gneezy. 2000. Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30: 163–182.
Erev, I., and A.E. Roth. 2001. On simple reinforcement learning models and reciprocation in the prisoner dilemma game. In *The adaptive toolbox*, ed. G. Gigerenzer and R. Selten. Cambridge: MIT Press.
Falk, A., E. Fehr, and U. Fischbacher. 2007. Testing theories of intentions – Fairness matters. *Games and Economic Behavior* (forthcoming).
Fehr, E., and K.M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114: 817–868.
Goeree, J.K., and C.A. Holt. 2001. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* 91: 1402–1422.
Henrich, J., et al. 2005. 'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28: 795–815.
Ho, T.H., C.F. Camerer, and K. Weigelt. 1998. Iterated dominance and iterated best response in experimental 'p-beauty contests'. *American Economic Review* 88: 947–969.
Ho, T.H., C.F. Camerer, and J.-K. Chong. 2007. Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory* 133: 177–198.
Hoffman, E., K. McCabe, K. Shachat, and V. Smith. 1994. Preferences, property-rights, and anonymity in bargaining games. *Games and Economic Behavior* 7: 346–380.
McKelvey, R.D., and T.R. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10: 6–38.
McKelvey, R.D., and T.R. Palfrey. 1998. Quantal response equilibria for extensive form games. *Experimental Economics* 1: 9–41.
McKelvey, R.D., and T.R. Palfrey. 2001. *Playing in the dark: Information, learning, and coordination in repeated games*. Princeton: Princeton University Press.
Nagel, R. 1995. Unraveling in guessing games: An experimental study. *American Economic Review* 85: 1313–1326.
Rabin, M. 1993. Incorporating fairness into game-theory and economics. *American Economic Review* 83: 1281–1302.
Roth, A.E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review* 81: 1068–1095.
Schlag, K.H., and G.B. Pollock. 1999. Social roles as an effective learning mechanism. *Rationality and Society* 11: 371–397.
Selten, R., and J. Buchta. 1999. Experimental sealed bid first price auctions with directly observed bid functions. In *Games and human behavior: Essays in honor of Amnon Rapoport*, ed. D.V. Budescu, I. Erev, and R. Zwick. Mahwah: Lawrence Erlbaum.
Stahl, D. 1999. *Sophisticated learning and learning sophistication*, Working paper. Austin: University of Texas.
Stahl, D.O. 2000. Rule learning in symmetric normal-form games: Theory and evidence. *Games and Economic Behavior* 32: 105–138.
Stahl, D.O., and P. Wilson. 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior* 10: 218–254.
Weibull, J. 1995. *Evolutionary game theory*. Cambridge, MA: MIT Press.

Behavioural Genetics

William T. Dickens

Abstract

The study of how variation in genetic endowment affects behaviour has shown that a surprisingly wide range of human activities are subject to substantial genetic influence. Studies of the covariance of traits in more and less distant relatives that take into account the impact of family environment have been the main method used to demonstrate this. This article provides a brief introduction to the mechanisms of heredity, and then a discussion of the methods used by behavioural geneticists and their limitations. Both the traditional variance decomposition methods and the newer molecular genetic methods are described and discussed.

Keywords

Adoption studies; Behavioural genetics; Chopstick problem; Cognitive ability; Environment vs heredity; Evolutionary psychology; Genome; Heredity; Heritability; Linkage studies; Twin studies

JEL Classifications

D1

While defining itself as the study of genetic influences on behaviour, behavioural genetics has been mainly concerned with demonstrating and quantifying the contribution of genetic variation to variation in human behavioural traits. As such, it contrasts with the related field of evolutionary psychology that attempts to understand how some behavioural traits common to all humans have been shaped by evolution.

The large and growing literature on the impact of genetic variation on behaviour leaves no room for doubt that genetic endowment is an important influence on a surprisingly wide range of behaviours. Behavioural genetics has relied mainly on the study of relatives with different degrees of relatedness or adoption to estimate the contributions of genetic variation and shared family environment to explaining crosssectional variation in behavioural characteristics. More recently, behavioural geneticists have been extending their methodology to use relational studies to examine the covariation of different behavioural traits, and molecular genetic methodologies to trace the sources and causes of genetically induced differences in behaviour.

Below I give a brief introduction to the mechanics of heredity. This is a necessary introduction to the methods of behavioural genetics, which I explain next.

Mechanics of Heredity

The human genetic code is contained in 23 pairs of chromosomes made up of deoxyribonucleic acid or DNA. A DNA molecule consists of two backbone strands that are held apart by molecular pairs of four bases. A sequence of these four chemicals along one of the backbone strands encodes the plans for the different proteins from which our bodies are made. Other parts of the code are thought to control when proteins are created and in what quantities. There are about three billion base pairs on just one set of 23 chromosomes. A sequence of base pairs that codes the information for a protein or some other function is called a ‘gene’.

Of the three billion base pairs all but about three million are the same in all humans. Where base pairs differ it is said that a polymorphism exists. When a gene contains one or more polymorphic base pairs there will be different versions of the gene. Different versions of the same gene are referred to as alleles.

A person’s genotype is determined by what alleles that person has, while the physiological characteristics or behaviours that geneticists study are referred to as the phenotype. Any given phenotypic behaviour can be the result of having a particular genotype, a particular environmental influence, or some combination of the two. Phenotypic traits are said to be qualitative if they take a limited number of discrete forms and quantitative if they vary continuously. So the presence of the symptoms of Huntington’s disease, a degenerative neurological disorder that affects older people, is a qualitative trait while one’s score on an IQ test is a quantitative trait.

Genetic influence on a phenotype can involve one or more genes. For example, people who have the allele for Huntington’s disease in the single gene encoding the huntingtin protein will contract it. Those who don’t won’t. Contrast that with the genetic influence on measured cognitive ability, which is thought to involve many genes, each of which has a very small effect on scores on tests of mental ability. When many genes influence a phenotypic trait, it is said to be polygenic.

Both qualitative and quantitative traits can be polygenic. A trivial example of a qualitative trait that is polygenic would be having an IQ score over 130. Other than some psychopathologies, most of the behaviours studied are thought to be polygenic with differences in each gene, making only a small contribution to differences in behaviour. In theory a quantitative trait could be influenced by a single gene that influenced the mean of the trait while environment determined the variance around the mean, but no examples of this have been identified.

Normally people inherit 46 chromosomes – 23 from their mothers and 23 from their fathers. Since there are many genes on any one chromosome, the inheritance of different traits can be linked if genes on the same chromosome influence

the traits. However, the linkage is not perfect. In the process of creating the chromosomes that will be passed on to one's children in gamete cells (ova and sperm), contiguous parts of each pair of chromosomes can be swapped so that the chromosome that is passed onto one's child is a combination of parts from both of one's parents. This happens on average about once per chromosome in humans. Thus, traits that are influenced by genes located close together on the same chromosome are more likely to be inherited together than genes on the same chromosome that are at distant loci. As will be described later, this fact can be used to identify the location of the genes that affect a particular trait.

If one has different alleles for the same gene on each of a pair of chromosomes there are different possible impacts. In some cases, certain alleles will always be expressed (influence phenotype) if they are present. Such alleles are termed 'dominant'. Other alleles for the same gene are called 'recessive' and will be expressed only if they are not paired with a dominant allele. In other cases, having two different alleles will have an effect on phenotype halfway between the effect of having two of the one allele and the effect of having two of the other. In this case genetic effects are termed 'linear and additive'.

There can be interactions between multiple genes in creating effects on phenotype. The phenomenon is called 'epistasis'. For example, there is epistasis if two different alleles of two different genes must be present for a phenotypic trait to be present. In this case, genetic effects on this trait will not be linear and additive.

Relational Studies

Arguably the first behavioural genetics study was Galton's *Hereditary Genius* (1869) in which he looked at patterns of career success in English families. He showed that close relatives of prominent men were also likely to achieve distinction, but that the probability fell with more and more distant relatives. While a genetic basis for ability would explain this pattern, so would family connections and a host of other environmental factors. Modern behavioural genetics research uses relational data, but in a way that attempts to control for family environment.

The simplest version of this type of study looks at the behavioural similarity of identical (or monozygote) twins who are raised apart. Such twins are genetic copies of each other as they grew from the same fertilized egg, but, if they are reared apart, then environmental similarities can't explain any behavioural similarities. If one assumes that genetic and environmental influences on a trait are linear and additive, then one can write

$$P = hG + cS + eN \tag{1}$$

where P is a measure of the phenotypic behaviour, G is genetic endowment, S is an index of the influence of shared family environment, and N is an index of the influence of environmental factors not shared by family members. The variables G, S and N are not observed, but the parameters h, c and e can still be estimates. If all variables are measured as standard deviations from their means, and G, S and N are uncorrelated, then h, c and e will be the correlations of the respective variable with P and their squares will be the fraction of variance in P that is explained by each. The fractions of variance in P explained by genetic endowment, shared family environment, and non-shared environment are commonly denoted h^2, c^2 and e^2. The sum of the squared coefficients will be one. Under the assumptions that the S's and N's of identical twins raised apart are uncorrelated, the expected correlation of P for pairs of twins is h^2 or the fraction of variation in the population explained by differences in genetic endowments. This statistic is referred to as the heritability of the trait P.

If one also has data on the correlation of the behaviour for identical twins raised together, one can construct an estimate of the fraction explained by the two environmental components as well. Under the assumption that identical twins raised together have both the same G and the same value for S, the correlation of P across pairs of identical twins raised together will be $h^2 + c^2$. So the difference between the correlation of P for identical

twins raised apart and those raised together will be the fraction of variance explained by shared family environment, and 1 minus that correlation will equal the share explained by non-shared environment.

With one additional assumption it is not necessary for the adopted siblings to be identical twins. Since natural siblings receive half of their genes from each parent and the genes received from each parent are in some sense a random subset of the parents' genes, it is not unreasonable to assume that the correlation of *G* for siblings who are not identical twins will be .5. In that case the expected correlation of a phenotype behaviour for siblings raised apart will be .5 h^2, and multiplying that value by 2 yields an estimate of the fraction of variance in the population explained by variation in genetic endowments. Once again, the difference between the correlation for siblings raised apart and those raised together will provide an estimate of the fraction of variance explained by shared family environment. The share attributable to non-shared environment can be computed as 1 minus the sum of the shares of genetic endowment and family environment.

If the effects of genetic endowment are not linear, then heritability estimates derived from studying twins adopted apart will be larger than those for siblings raised separately. Since monozygote twins are genetically identical, they will be affected by dominant genes and interaction effects between genes (epistasis) in exactly the same way. Thus, studies of identical twins measure what is called 'broad- sense heritability' (denoted H^2) unless dominance and epistasis effects are absent. In the presence of dominance and epistasis effects the correlation of phenotypes between normal sibling pairs raised apart will be less than half of that of identical twins raised apart. Twice the correlation for normal siblings raised apart is said to measure narrow-sense heritability since it doesn't reflect the contribution of nonlinear genetic effects.

Estimated variance shares from adoption studies can be criticized on a number of grounds. Siblings raised apart, and particularly twins, will share aspects of their prenatal environment at least. They may also share their post-natal environment if they are not adopted away immediately. Also, siblings who are put up for adoption may end up in similar environments for a number of reasons. They may be adopted by relatives, or they may be adopted through the same agency that places children with parents of a particular social class in a particular geographic area. Adopting families may be matched to the socio-economic status of the biological mother. Similar environments will cause adoptees to resemble each other even if there is no effect of genetic endowment and will bias estimates of heritability upward. Adoption itself may affect the trait, leading to an overestimate of heritability and an underestimate of the role of shared environment.

Even if adoption doesn't place siblings in similar environments, it almost certainly restricts the range of environments compared with those occupied by children living with their natural parents, as adoption agencies rigorously screen parents wishing to adopt. Stoolmiller (1999) argues that this restriction of range leads adoption studies to underestimate the role of shared family environment and overestimate the importance of genetic differences in explaining variance in the general population, since there is much more variation in family environment in the general population than in adopting families. This illustrates an important characteristic of heritability estimates – they apply only to the population in which they are estimated. Populations with different amounts of variation in environment or genetic endowment would exhibit different heritabilities. Finally, the assumption that the correlation of normal siblings with no environment in common will be exactly .5 h^2 is probably wrong for another reason. It assumes that each parent's genes for a trait are a random draw from the population – that is, that men and women don't choose each other as mates on the basis of the characteristic being studied or anything related to it. If parents are likely to have genes for the trait in common, then the expected correlation will be higher and multiplying it by 2 will overestimate heritability. If opposites attract, then multiplying the sibling correlation by 2 will understate heritability. Estimates of the variance explained by shared family environment

will be affected and biased in the opposite direction to heritability.

An alternative to adoption studies are those that contrast the similarity of identical twins with that of fraternal twins. Identical twins are genetic copies of each other while fraternal twins are no more alike genetically than brothers and sisters. Thus we would expect identical twins to be more similar for traits that are subject to genetic influence. Again, under the standard assumptions, the correlation of identical twins in a population will be $h^2 + c^2$. If one assumes that fraternal twins' genetic endowments have a correlation of .5, then their correlation will be .5 $h^2 + c^2$. Thus, twice the difference between the correlation for identical and fraternal twins is an estimate of heritability. The fraction of variance explained by shared environment will be equal to the identical twin correlation minus the estimate of heritability, and that of non-shared environment will equal 1 minus the identical twin correlation.

Twin studies, too, can be criticized on a number of grounds. The assumption that the correlation of genetic endowment for fraternal twins will be .5 rests on random mating. If husbands and wives tend to have similar genetic endowments for the characteristic being studied, then the fraternal twin correlation will be greater than .5, and doubling the difference between fraternal and identical twins will understate heritability and overstate the role of shared environment. On the other hand, if there are dominance and epistasis effects, doubling the difference will overstate both broad and narrow sense heritability.

A common criticism of twin studies is that identical twins have more similar environments than fraternal twins and that accounts for some of their greater similarity. Whether or not this is a valid criticism, it certainly illustrates a common misunderstanding about the meaning of heritability. If identical twins have more similar environments because they behave in more similar ways and create for themselves more similar environments, some would say that it is legitimate to attribute the influence of environment of this sort to genetic endowment. In the same sense, natural siblings may have more similar environments than adopted siblings – even if they are raised apart – because their more similar genes induce more similar behaviour which induces more similar responses from their environment. If two siblings are both genetically predisposed to be taller, they may both end up playing on the high-school basketball team, where they receive professional coaching which greatly improves their skills. The similarity of their basketball skill is a direct effect of similar environments, but it is also an indirect effect of genetic endowment. Both twin and adoption studies will attribute such induced environmental effects to genetic endowment.

A common error in the interpretation of heritability estimates is the assumption that, if heritability is high, the effects of environment must be small and the trait not easy to change through environmental intervention. However, if heritability estimates attribute to genetic endowment indirect effects that come through environment, it's easy to see that this is not the case (see the discussion of malleability in the entry on cognitive ability). If a tall person is good at basketball mainly because he has received good coaching, then the skill of shorter people can probably be improved a great deal by coaching as well (even if they can never be quite as good as the tall person). When genetic endowment has both direct physiological effects on a trait and indirect effects through induced environment, there is gene x environment correlation. Relaxing the assumption that genetic endowment and environmental influences are correlated doesn't invalidate heritability estimates, but it does change their interpretation as just explained. The fractions of variance explained by shared and non-shared environment in twin and adoption studies are not the full effect of environment, but the fractions explained by the residual environment – that part that can't itself be explained by differences in genetic endowment.

There is another reason why high heritability estimates do not mean that the effects of environment are necessarily weak. Recall that heritability estimates are valid only in the population in which they are estimated. If we were to study nearsightedness in a population of people who were not wearing corrective lenses, we would find it highly heritable. If we studied scores on an eye test allowing people to wear their corrective lenses,

we would probably find very low heritability of test scores. The high heritability of nearsightedness in the first case certainly wouldn't mean that we couldn't treat it with corrective lenses.

Interaction of environment and genotype can create problems of interpretation similar to the just-described problems caused by the correlation between genotype and environment. Interaction is said to exist when environment has different effects depending on a person's genotype. In this case genetic effects are not linear and additive and the variance shares computed using standard behavioural genetic methods do not provide a meaningful measure of effects of genetic endowment and environment on the trait. None the less, high estimates of heritability for a population still indicate a substantial role for genetic variation in causing variation in the trait.

Some of the shortcomings of twin studies and adoption studies can be overcome by combining data from the two. Since they are subject to different biases, if results for the two types of studies are very similar, one can have some confidence that the biases are not important. Data from the two types of studies can be formally combined and used to estimate more elaborate models of inheritance that relax one or more assumptions such as linearity, random mating, or similar treatment of identical and fraternal twins. Information on other types of relations and more distant relations can be added to model building studies as well.

Of all the behaviours to which relational methods have been applied, the one that has received the most attention is scores on tests of cognitive ability. These studies have been extremely controversial – at least in part because of the widespread misunderstanding that high heritability precluded an important role for environment. Today it is widely accepted that the heritability of cognitive test scores in adults is very high (0.6 or more; Neisser et al. 1996; Plomin et al. 2000, pp. 164–77), but it is understood that this does not imply a limited role for environment (as genetic endowment may be acting indirectly through the environment).

Besides cognitive ability, a wide range of other behaviours have been studied. The degree to which people display the symptoms of a number of psychopathologies has been shown to be subject to genetic influence (Plomin et al. 2000, chs 8 and 12). Major measurable aspects of personality (Loehlin 1992), religiosity (Waller et al. 1990), attitudes towards one's job (Lykken et al. 1993), social attitudes (Martin et al. 1986) (including political conservatism; Eaves et al. 1997), education (Behrman and Taubman 1989), earnings (Taubman 1976), and even the amount of time spent watching television (Plomin et al. 1990), have all been shown to be subject to genetic influence. In most cases, studies find that the fraction of variance explained by variation in genetic endowment is large and greater than the fraction explained by family environment (Turkheimer 2000). Also interesting are the exceptions that have been found to this general pattern. For example, how often one attends church is influenced by one's genetic endowment, but not the type of church one attends.

A relatively recent development in relational studies is their use to analyse the sources of covariance between different measures of behaviour. By using similar assumptions to those used to identify variance shares, it is possible to tell whether correlations between variables are due mainly to common genetic factors, common environmental factors or both. For example, tests of cognitive ability are strongly correlated with scores on achievement tests and both are highly heritable. Are the same genetic factors responsible for both (as would be the case if genetic influence on achievement came entirely through its effects on cognitive ability)? For the most part they are, though some genetic influence is specific to achievement (Plomin et al. 2000, p. 201).

Animal Models and Molecular Genetics Studies

Work with animals allows behavioural geneticists to do many things that are impossible with human subjects. For example, animals can be bred for certain behavioural traits and then the specially bred animals can be used in experiments. One of the most interesting demonstrations of gene

x environment interaction comes from a study of two strains of rats that had been bred for their performance in solving mazes (Cooper and Zubek 1958). One strain was bred for superior performance and one for inferior performance. Rats raised in very sparse environments performed poorly in solving mazes no matter what their genetic endowment. Rats raised in enriched environments performed much better and there was little effect from their genetic endowment. However, rats raised in normal laboratory environments showed large differences consistent with their genetic endowments.

Animal studies can be particularly useful when combined with some of the new molecular genetic techniques. Certain genes can be turned off and the impact on behaviour studied. Genetic mutations can be created in experimental animals and the impact of the mutation on behaviour examined. Selectively bred animals can be compared for the frequency of different alleles to determine where genes that influence a trait are located.

Searches of this sort are facilitated by the previously described tendency for genes that are located close together on a chromosome to be inherited together. Suppose, for example, that animals that had been bred for an extreme form of some behaviour showed a much higher frequency of one allele on one chromosome than did the population from which they were bred. This would not mean that that allele played a role in the development of that trait, but it would make it more likely than not that one or more genes on the chromosome on which the gene was located played some role. The allele that is found to be associated with the trait being studied is said to be a marker for the trait, while the genes with the polymorphisms that matter for the trait are said to be trait loci. If the trait is a quantitative trait, each locus is referred to as a quantitative trait locus (QTL).

If several markers are studied on the same chromosome, some may be found to be more highly associated with the trait than others. The more highly associated markers are likely to be closer to one or more trait loci since, the closer two genes are together on the same chromosome the more likely it is that they will be inherited together.

This technique has been used to identify the location of genes with a large role in determining differences in fearfulness in mice. The same sequence of genes exists in the human genome and it is possible that variations in them may explain why some people develop anxiety disorders and some don't. Understanding the role of these genes may lead to more effective treatment.

Association techniques can also be used in humans, but are subject to a number of problems. In the example just discussed, the mice studied were all bred from the same homogenous population. The breeding for the trait is likely to have induced any association found between a marker and a phenotype trait. However, in human populations markers and traits could be associated even if there was no genetic influence on the behaviour. This is referred to as the 'chopstick' problem, which is named after a commonly cited example of a spurious association. In a population that included native Chinese and Europeans, using chopsticks would be associated with any marker more common in Chinese. This problem can be partially overcome by studying more homogenous populations or contrasting sibling pairs, as differences in marker frequency are more likely to signal genetic causation in these cases. In the extreme, studies can be done on large extended families. The families can be studied for co-transmission of the trait and particular alleles. These are termed 'linkage studies'. Linkage studies were used to identify the gene responsible for Huntington's disease.

Linkage studies solve another problem of association studies in humans. Within a family, even markers fairly distant from a trait locus will have some degree of association with the trait. In the general population, markers are likely to be associated with traits only if they are trait loci themselves or are located very close to them, as recombination of chromosomes will eventually break down the association of any marker that is not a trait locus with the trait after a sufficient number of generations. A much smaller number of markers can be used to scan for the location of

trait loci in a linkage study than in a study looking for association in the general population. However, linkage studies are not very good at finding QTLs when there are many genes contributing to a phenotype. Association studies in large populations are more promising, but only if the area of the genome to be examined can be narrowed on the basis of hypothesis about what systems might be involved. So far this approach has shown some promise. For example, associations have been found between a particular allele for a dopamine receptor gene and hyperactivity disorder in children (Thapar et al. 1999).

The Future

Relational studies have demonstrated that variation in a surprisingly wide range of behaviours is substantially influenced by genetic differences. Molecular genetics has begun to discover some of the mechanisms by which genetic differences cause differences in behaviour, but work of this sort has barely scratched the surface, and further development faces some difficult obstacles. Most of the behaviours that have been studied are thought to be affected by many different genes, each of which has a small effect. This will make identifying QTLs difficult without some theory of what physiological processes might be involved and where the genes affecting those processes are in the genome. But what theory might one have about the location of physiological processes affecting, for example, time spent watching television? When one begins to think about the many ways in which physiological differences could affect a wide range of behaviours, the task seems daunting. Suppose there was an allele that when present made people feel more discomfort when they were cold than others without the allele. Such people might be inclined to spend more time inside watching TV. They might also be less athletic and/or more likely to spend a lot of time reading. If they read more, they might have larger vocabularies and score better on IQ tests. If their reading made them more sceptical, they might be less likely to attend church. Depending on how myriad and diffuse such cascading effects are, it might be impossible to understand how more than a small fraction of genetically induced differences in behaviour comes about. Still, that doesn't mean that valuable knowledge can't be gained from studying the pathways that can be identified. Such knowledge might accumulate faster if those studying the genetic influences on behaviour concentrated less on refining estimates of heritability and more on analysing the role of genetic differences in explaining the covariance of different behaviours.

See Also

▶ Cognitive Ability

Bibliography

Behrman, J.R., and P. Taubman. 1989. Is schooling mostly in the genes? *Journal of Political Economy* 97: 1425–1446.

Cooper, R.M., and J.P. Zubek. 1958. Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Canadian Journal of Psychology* 12: 159–164.

Eaves, L.J., J.L. Silberg, J.M. Meyer, H.H. Maes, E. Simonoff, A. Pickles, M. Rutter, M.C. Neale, C.A. Reynolds, M.T. Erikson, et al. 1997. Genetics and developmental psychology: 2 The main effects of genes and environment on behavioral problems in the Virginia twin study of adolescent behavioral development. *Journal of Child Psychology and Psychiatry* 38: 965–980.

Galton, F. 1869. *Hereditary genius: An enquiry into its laws and consequences*. London: Macmillan.

Loehlin, J.C. 1992. *Genes and environment in personality development*. Newbury Park: Sage Publications, Inc.

Lykken, D.T., T.J. Bouchard, M. McGue, and A. Tellegen. 1993. Heritability of interests: A twin study. *Journal of Applied Psychology* 78: 649–661.

Martin, N.G., L.J. Eaves, A.C. Heath, R. Jardine, L.M. Feingold, and H.J. Eysenck. 1986. Transmission of social attitudes. *Proceedings of the National Academy of Science* 83: 4364–4368.

Neisser, U., G. Boodoo, T.J. Bouchard Jr., A.W. Boykin, N. Brody, S.J. Ceci, D.F. Halpern, J.C. Loehlin, R. Perloff, R.J. Sternberg, and S. Urbina. 1996. Intelligence: Knowns and unknowns. *American Psychologist* 51: 77–101.

Plomin, R., R. Corley, J.C. DeFries, and D.W. Fulker. 1990. Individual differences in television viewing in early childhood: Nature as well as nurture. *Psychological Science* 1: 371–377.

Plomin, R., J.C. DeFries, G.E. McClearn, and P. McGuffin. 2000. *Behavioral genetics*, 4th ed. New York: Worth.

Stoolmiller, M. 1999. Implications of the restricted range of family environments for estimates of heritability and non-shared environment in behavior-genetic adoption studies. *Psychological Bulletin* 125: 392–409.

Taubman, P. 1976. The determinants of earnings: Genetics, family and other Environments: A study of male twins. *American Economic Review* 66: 858–870.

Thapar, A., J. Holmes, K. Poulton, and R. Harrington. 1999. Genetic basis of attention deficit and hyperactivity. *British Journal of Psychiatry* 174: 105–111.

Turkheimer, E. 2000. Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science* 9(5): 160–164.

Waller, N.G., B.A. Kojetin, T.J. Bouchard, D.T. Lykken, and A. Tellegen. 1990. Genetic and environmental influences on religious interests, attitudes, and values: A study of twins reared apart and together. *Psychological Science* 1: 138–142.

Behavioural Public Economics

B. Douglas Bernheim and Antonio Rangel

Abstract

Behavioural public economics incorporates ideas from behavioural economics, psychology, and neuroscience in the analysis and design of public policies. This article provides an introduction to its methods and discusses its application to savings and addiction policy.

Keywords

Addiction; Behavioural public economics; Budget constraints; Compulsory saving; Default options; Imperfect decision processes; Intertemporal choice; Lump-sum taxes; Myopia; Neoclassical public economics; Neuroscience; Pigouvian taxes; Psychology and economics; Tax incentives for saving; Well-being

JEL Classifications

H3

Interest in the field of psychology and economics has grown in recent years, stimulated largely by accumulating evidence that the neoclassical model of consumer decision-making provides an inadequate description of human behaviour in many economic situations. Scholars have begun to propose alternative models that incorporate insights from psychology and neuroscience. Some of the pertinent literature focuses on behaviours commonly considered 'dysfunctional', such as addiction, obesity, risky sexual behaviour, and crime. However, there is also considerable interest in alternative approaches to more standard economic problems such as saving, investing, labour supply, risk-taking, and charitable contributions.

Behavioural public economics (BPE) is the label used to describe a rapidly growing literature that uses this new class of models to study the impact of public policies on behaviour and well-being (see Bernheim and Rangel 2006a, for a more comprehensive review).

Background: The Neoclassical Approach to Public Economics

Public economic analysis requires us to formulate models of human decision-making with two components – one describing choices, and the other describing well-being. Using the first component, we can forecast the effects of policy reforms on individuals' actions, as well as on prices and allocations. Using the second component, we can determine whether these changes benefit consumers or harm them.

The neoclassical approach assumes that individuals' choices can be described *as if* generated by the maximization of a well-defined and stable utility function subject to feasibility and informational constraints. Neoclassical welfare analysis proceeds from the premise that, when evaluating policies, the government should act as each individual's proxy, extrapolating his preferred choices from observed decisions in related situations. This premise justifies the use of the *as-if* utility function as a gauge of well-being. In effect, this approach uses the same model for positive and normative analysis.

Within the neoclassical paradigm, government policy can affect behaviour and welfare only if it changes the decision maker's information or budget constraint. For example, vaccination campaigns may influence behaviour by providing information concerning the risks of a disease and the advantages of taking preventive action, while cigarette taxes may alter choices by raising the cost of smoking.

From the neoclassical perspective, government intervention in private markets is justified to enforce property rights, correct market failures, and address inequity by redistributing resources. Standard examples of interventions motivated by market failures include the use of taxes and subsidies to correct externalities, the provision of public goods, and the introduction of social insurance when private risk sharing is inefficient.

The accomplishments of neoclassical public economics, such as the theories of optimal income taxation and corrective environmental policy, are considerable. However, there is growing concern that this paradigm does not adequately address a number of important public policy challenges – for example, what to do about 'self-destructive' behaviours such as substance abuse, or about the apparently myopic choices of those who save 'too little' for retirement. Since the neoclassical welfare criterion respects all voluntary consumer choices (conditional on the information in the consumer's possession), it rules out the possibility of enhancing well-being by correcting 'poor' choices (except through the provision of information).

The Behavioural Approach to Public Economics

A key feature of BPE is the potential divergence of positive and normative models. Even when it is assumed that individuals are endowed with well-behaved lifetime preferences, decision processes may translate these preferences to choices imperfectly. To conduct positive analysis, one employs a model of the potentially imperfect decision process. To conduct normative analysis, one uses a well-defined welfare relation. In stark contrast to the neoclassical approach, the welfare relation may prescribe an alternative other than the one that the individual would choose for himself, at least under some conditions.

The analysis of addiction presented in Bernheim and Rangel (2004) illustrates this approach. Our model assumes that people attempt to optimize given their preferences, but randomly encounter conditions that trigger systematic mistakes, the likelihood of which evolves with previous substance use. The model is based on the following three premises. First, use among addicts is sometimes a mistake and sometimes rational. Second, experience with an addictive substance sensitizes an individual to environmental cues that trigger mistaken usage. Third, addicts understand their susceptibility to cue-triggered mistakes and attempt to manage the process with some degree of sophistication. The first two premises are justified by a body of research in psychology and neuroscience, which shows that, after repeated exposure to an addictive substance, the brain tends to overestimate the hedonic consequences of drug consumption upon encountering environmental cues that are associated with past use. The third premise is justified by behavioural evidence indicating that users are often surprisingly sophisticated and forward looking.

The (β,δ)-model of intertemporal choice (Strotz 1956; Phelps and Pollack 1968; Laibson 1997; O'Donoghue and Rabin 1999b, 2001) also illustrates the BPE approach. Psychologists have found that people often act as if they attach disproportionate importance to immediate rewards relative to future rewards, especially in situations where cognitive systems are overloaded. (For a recent review of this literature, see Frederick et al. 2002; Loewenstein et al. 2003.) To capture this tendency, the (β,δ)-model assumes that, in each period t, individuals behave as if they maximize a utility function of the form

$$u(c_t) + \beta\left[\sum_{k=t+1}^{T} \delta^{k-t} u(c_k)\right],$$

where $0 < \beta < 1$. In this framework, the parameter β represents the degree of *present bias* or *myopia*. The neoclassical model corresponds to the special

case where $\beta = 1$. With $\beta < 1$, behaviour is dynamically inconsistent. This complicates positive analysis, since behaviour no longer corresponds to the solution of single utility maximization problem.

Many analysts interpret present bias as a mistake. They argue that the individual's underlying well-being actually corresponds to the preferences revealed through choices that do not involve immediate rewards:

$$U(c_1, \ldots, c_T) = \sum_{t=0}^{T} \delta^t u(c_t).$$

Under this interpretation, $\beta < 1$ creates a tendency to consume excessively in the present.

These examples illustrate some important conceptual and methodological aspects of BPE. First, with behaviour and welfare modelled separately, BPE allows for the possibility of mistakes. In contrast to a neoclassical analyst, a BPE analyst can pose questions that presuppose possible divergences between behaviour and preferences, such as whether Americans save too little for retirement, or whether addicts engage in self-destructive behaviour. Within the BPE framework, one can test the hypothesis that individuals maximize their well-being, and measure the magnitude of their errors. Second, to justify either a positive representation of choice or a particular welfare criterion, a BPE analyst relies on evidence from psychology and neuroscience. This evidence can help economists pin down underlying preferences by identifying the mechanisms responsible for the decision-making errors. Good structural models of decision-making processes may also improve the quality of out-of-sample behavioural predictions, which are often required for policy evaluation.

Behavioural Policy Analysis

BPE models are extensions of neoclassical models. Thus, they imply that public policy can modify behaviour by changing budget constraints and/or information. For example, cigarette prices affect cigarette consumption in the Bernheim–Rangel addiction model, and savings are responsive to interest rates in most specifications of the (β,δ)-model.

In addition, the BPE framework introduces new channels through which public policy can affect behaviour and welfare. In particular, it allows for the possibility that some public policies can influence behaviour *directly* by activating particular cognitive processes, even when they leave budget constraints and information unchanged.

For example, Brazil and Canada require every pack of cigarettes to display a prominent, viscerally charged image depicting some deleterious consequences of smoking, such as lung disease and neonatal morbidity. Since the consequences of smoking are well known, this policy has no effect in information or budget constraints. And yet the Bernheim–Rangel theory of addiction allows for the possibility that a sufficiently strong counter-cue could reduce the probability of a mistake by triggering thought processes that induce users to resist cravings. When successful, this policy affects behaviour by activating particular cognitive processes.

Another striking example involves the effects of default options in employeedirected pension plans. A 'default option' is the outcome resulting from inaction. For a neoclassical consumer, choices depend only on preferences, information, and constraints. Consequently, in the absence of significant transaction costs, default options should be inconsequential. However, in the context of decisions concerning saving and investment, defaults seem to matter a great deal. For example, with respect to 401(k) plans (employer-sponsored retirement savings accounts in the United States that receive preferential tax treatment), there is considerable evidence that default options affect participation rates, contribution rates, and portfolios (Madrian and Shea 2001; Choi et al. 2004). Yet, arguably, a default neither affects opportunities (since transaction costs are low) nor provides new information.

While BPE models admit traditional justifications for government intervention in private markets (the enforcement of property rights, the correction market failures, and the redistribution of resources), they also introduce novel justifications. For example, public policy may improve

welfare by reducing the size, likelihood, or consequences of mistakes. As shown in the next two sections, this can lead to conclusions that are strikingly at odds with those generated by the neoclassical model.

Example: Addiction Policy

In the neoclassical theory of rational addiction (Becker and Murphy 1988), government intervention may be justified *only* when it corrects market failures involving addictive substances, such as second-hand smoking, or when it combats ignorance or misinformation. In contrast, in our model of addiction (Bernheim and Rangel 2004), government intervention may also be justified when it reduces the frequency, magnitude, and consequences of mistakes. These considerations give rise to a number of non-standard policy implications.

Limitations of informational policy In practice, public education campaigns (such as anti-smoking and anti-drug initiatives) have achieved mixed results. Our view of addiction highlights a fundamental limitation of informational policy: contrary to standard theory, one cannot assume that even a highly knowledgeable addict always makes informed choices. Information about the consequences of substance abuse may affect initial experimentation with drugs, but cannot alter the neurological mechanisms through which addictive substances subvert deliberative decision-making.

Beneficial harm reduction If addiction results from randomly occurring mistakes, various interventions can serve social insurance objectives by ameliorating some of its worst consequences. For instance, subsidization of rehabilitation centres and treatment programmes (particularly for the indigent) can moderate the financial impact of addiction and promote recovery. Likewise, the free distribution of clean needles can moderate the incidence of diseases among heroin addicts. In some cases, it may even be beneficial to make substances available to severe addicts at low cost, a policy used in some European countries.

Counterproductive disincentives Policies such as 'sin taxes' strive to discourage use by making substances costly. This is potentially justifiable on the grounds that use generates negative externalities. Even higher taxes (whether implicit or explicit) might be justified if they also reduce 'unwanted' use. Unfortunately, the compulsive use of addictive substances is probably much less sensitive to costs and consequences than is deliberative use. Consequently, imposing costs on users in excess of the standard Pigouvian levy will likely distort deliberate choices detrimentally, without significantly reducing problematic compulsive usage. In addition, policies that impose high costs on use may thwart social insurance objectives by exacerbating the consequences of uninsurable risks associated with the use of addictive substances, such as poverty and prostitution. Accordingly, for some substances the optimal rate of taxation for addictive substances may be significant *lower* than that the standard Pigouvian levy (see Bernheim and Rangel 2005, for simulation results).

Policies affecting cues Since environmental cues appear to trigger addictive behaviours, public policy can also influence use by changing the cues that people normally encounter. One approach involves the elimination of problematic cues. For example, advertising and marketing restrictions of the type imposed on sellers of tobacco and alcohol suppress one possible artificial trigger for compulsive use. Since one person's decision to smoke may trigger another, confining use to designated areas may reduce unintended use. Another approach involves the creation of counter-cues, which we discussed above. Policies that eliminate problematic cues or promote counter-cues are potentially beneficial because they combat compulsive use while imposing minimal inconvenience and restrictions on rational users.

Facilitation of self-control Most behavioural theories of addiction potentially justify policies that provide better opportunities for self-regulation

without making particular choices compulsory. In principle, this helps those who are vulnerable to compulsive use without encroaching on the freedoms of those who would deliberately choose to use. Laws that limit the sale of a substance to particular times, places, and circumstances may facilitate self-regulation. Well-designed policies could in principle accomplish this objective more effectively. For example, a number of states have enacted laws allowing problem gamblers to voluntarily ban themselves from casinos. Alternatively, if a substance is available only by prescription, and if prescription orders are filled on a 'next day' basis, then deliberate forward-looking planning becomes a prerequisite for availability. In the absence of a pervasive black market, recovering heroin addicts could self-regulate problematic compulsive use by carefully choosing when, and when not, to file requests for refills.

Example: Savings Policy

The (β,δ)-model of savings also exemplifies the novel policy insights generated by the BPE approach. For example, this model implies that many individuals will save too little for retirement, and that there may be Pareto improving policy interventions even in the absence of capital market distortions – a conclusion that is at odds with the neoclassical framework. Other notable implications include the following:

Mandatory savings policies Within the (β,δ) framework, compulsory saving may be welfare-enhancing if it fully crowds out private saving (in the form of liquid assets) at some point during the life cycle (Imrohoroglu et al. 2003; Diamond and Koszegi 2003). This provides a rationale for mandatory savings programmes, which are pervasive across the world, and which are more difficult to justify within the neoclassical framework.

Saving subsidies On the assumption that (*a*) the population includes some individuals with self-control problems and (*b*) the social welfare function is continuous and concave, a small subsidy for saving financed with lump-sum taxes is welfare improving (O'Donoghue and Rabin 2006; Krusell et al. 2000, 2002). Intuitively, the subsidy produces a first-order improvement in the well-being of individuals with self-control problems (since they save too little), and only a second-order reduction in the well-being of those without self-control problems. This provides a possible rationale for tax-favoured savings programmes, such as, in the United States, 401(k) plans and Individual Retirement Accounts (IRAs).

Credit restrictions Introducing restrictions on the availability of credit, for example, by regulating the distribution of revolving credit lines and mandating credit ceilings, can potentially enhance the well-being of those with self-control problems. For example, Laibson, Repetto and Tobacman (2004) estimate that the representative (β,δ) consumer would be willing to pay $2000 at the age of 20 to exclude himself from the credit card market.

Behavioural Public Economics Circa 2006

As of 2006, the rapidly growing field of BPE has demonstrated its value by enhancing our understanding of public policy in several areas, including savings and addiction. Nevertheless, the literature is still in its infancy. As time passes, we anticipate that the methods and tools of BPE will contribute new insights in these areas, as well as to other difficult public policy issues involving poverty, crime, corruption, violence, obesity, and charitable giving, among others.

In addition to providing new insights concerning the effects of familiar policies, research in BPE can also guide the design of new policies. One obvious goal is to reduce the frequency of mistakes among those who behave suboptimally without interfering with the choices of those who behave optimally. Some recent fieldwork by Thaler and Bernartzi (2004), who advocate a savings programme called Save More Tomorrow, illustrates the potential value of this approach. In this programme, a worker can allocate a portion of her future salary increases

towards retirement savings. Subsequently, she is allowed to change this allocation at a negligible transaction cost. In practice, 78 per cent of those who were eligible for the plan chose to participate, 80 per cent of participants remained in the plan through the fourth pay raise, and the average contribution rate for programme participants increased from 3.5 per cent to 13.6 per cent over the course of 40 months.

To date, progress in BPE has been somewhat hampered by the absence of a general framework for behavioural welfare analysis. Analysts tend to devise and justify welfare criteria on a case-by-case basis, rather than through the application of general principles. Ongoing research aims to fill this gap (see Bernheim and Rangel 2006b).

See Also

- ▶ Addiction
- ▶ Behavioural Game Theory
- ▶ Charitable Giving
- ▶ Neuroeconomics
- ▶ Public Goods Experiments

Bibliography

Becker, G., and K. Murphy. 1988. A theory of rational addiction. *Journal of Political Economy* 96: 675–700.

Bernheim, B.D., and A. Rangel. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94: 1558–1590.

Bernheim, B.D., and A. Rangel. 2005. From neuroscience to public policy: A new economic view of addiction. *Swedish Economic Policy Review* 12: 11–46.

Bernheim, B.D., and A. Rangel. 2006a. Behavioral public economics: Welfare and policy analysis with fallible decision-makers. In *Economic institutions and behavioral economics*, ed. P. Diamond and H. Vartiainen. Princeton: Princeton University Press (forthcoming).

Bernheim, B.D., and Rangel A. 2006b. Toward choice-theoretic foundations for behavioral welfare economics. *American economic review papers and proceedings*, (forthcoming).

Choi, J., D. Laibson, and B. Madrian. 2004. Plan design and 401(k) savings outcomes. *National Tax Journal* 57: 275–298.

Diamond, P., and B. Koszegi. 2003. Quasi-hyperbolic discounting and retirement. *Journal of Public Economics* 87: 1839–1872.

Frederick, S., G. Loewenstein, and T. O'Donoghue. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40: 351–401.

Imrohoroglu, S., A. Imrohoroglu, and D. Joines. 2003. Time inconsistent preferences and social security. *Quarterly Journal of Economics* 118: 745–784.

Krusell, P., B. Kuruscu, and A. Smith. 2000. Tax policy with quasi-geometric discounting. *International Economic Journal* 14(3): 1–40.

Krusell, P., B. Kuruscu, and A. Smith. 2002. Equilibrium welfare and government policy with quasi-geometric discounting. *Journal of Economic Theory* 105: 42–72.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112: 443–477.

Laibson D, Repetto A, and Tobacman J. 2004. *Estimating discount functions from lifecycle consumption choices. Working paper*. Harvard University.

Loewenstein, G., D. Read, and R. Baumister, ed. 2003. *Time and decision: Economic and psychological perspectives on intertemporal choice*. New York: Russell Sage Foundation.

Madrian, B., and D. Shea. 2001. The power of suggestion: inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics* 116: 1149–1187.

O'Donoghue, T., and M. Rabin. 1999b. Doing it now or later. *American Economic Review* 89: 103–124.

O'Donoghue, T., and M. Rabin. 2001. Choice and procrastination. *Quarterly Journal of Economics* 116: 121–160.

O'Donoghue, T., and M. Rabin. 2006. Optimal sin taxes. *Journal of Public Economics* 90: 1825–1849.

Phelps, E., and R. Pollack. 1968. On second-best national savings and game equilibrium growth. *Review of Economic Studies* 35: 185–199.

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

Thaler, R.H., and S. Bernartzi. 2004. Save more for tomorrow: Using behavioral economics to increase employee savings. *Journal of Political Economy* 112: S164–S187.

Bellman Equation

Yongseok Shin

Keywords

Bellman equation; Consumption smoothing; Convergence; Dynamic programming; Markov processes; Neoclassical growth theory; Value function

JEL Classifications
C61

Dynamic programming is a method that solves a complicated multi-stage decision problem by first transforming it into a sequence of simpler problems. Bellman equations, named after the creator of dynamic programming Richard E. Bellman (1920–1984), are functional equations that embody this transformation.

Take, for example, a typical maximization problem in economics:

$$\max_{\{u_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t F(x_t, u_t), \tag{1}$$

s.t. $x_{t+1} = g(x_t, u_t)$ and $u_t \in \Gamma(x_t)$, with x_0 given.

The set $\Gamma(x_t)$ consists of admissible values of the control variable u_t given the state variable x_t. We assume that $\Gamma(x_t)$ is non-empty for all x_t. We also assume that $F(x_t, u_t)$ is concave and that the set $\{(x_t, x_{t+1}): x_{t+1} = g(x_t, u_t), u_t \in \Gamma(x_t)\}$ is compact and convex. It is further assumed that $\beta \in (0, 1)$. This so-called sequence problem has an infinite number of controls $\{u_t\}_{t=0}^{\infty}$, and is generally intractable as it is. Dynamic programming reduces this infinite-dimensional problem into an infinite sequence of one-dimensional problems:

$$\max_{u \in \Gamma(x)} F(x, u) + \beta V(x'), \tag{2}$$

s.t. $x' = g(x, u)$.

The unknown function $V(x)$ represents the maximized value of the original problem starting from an arbitrary initial condition x, and is called the value function. In particular, $V(x_0)$ must be equal to the maximized value of the objective function in the original problem (1). Once $V(x)$ is known, the maximizer of (2) would take the form of an optimal decision rule, or a policy function: $u^* = h(x)$. Let the maximizer of the original problem (1) be $\{u_t^*\}_{t=0}^{\infty}$. Then $\{u_t^*\}_{t=0}^{\infty}$ can be generated from (2) recursively by $u^* = h(x_t)$ and $x_{t+1} = g(x_t, u_t^*)$, starting from the given x_0. Bellman called this connection between the sequence problem (1) and the recursive problem (2), the principle of optimality.

Now we have to solve for $V(x)$ and, subsequently, $h(x)$. To this end, we re-write (2) as follows:

$$V(x) = \max_{u \in \Gamma(x)} F(x, u) + \beta V(g(x, u)), \tag{3}$$

This functional equation in $V(x)$ is the Bellman equation. From the definition of $h(x)$, it follows that $V(x) = F(x, h(x)) + \beta V(g(x, h(x)))$.

Typically, the Bellman equation can be solved for the unknown $V(x)$ by value function iteration. This method can be described as follows.

1. Guess an arbitrary function $V_j(x), j = 0$.
2. Given $V_j(x)$, compute $V_{j+1}(x) = \max_{u \in \Gamma(x)} F(x, u) + \beta V_j(g(x, u))$.
3. Repeat Step 2 until the sequence of functions $\{V_j\}_{j=0}^{\infty}$ thus constructed converge. The limit of this sequence is the solution to the functional equation (3), $V(x)$.

Under some conditions (for example, Blackwell's sufficient conditions), it is proven that value function iteration recovers the unique solution to (3) starting from an arbitrary initial guess $V_0(x)$. See Bertsekas (1976) or Stokey and Lucas (1989) for detailed expositions on convergence. The procedure may sound straightforward, but, in practice, it is impossible (with few exceptions) to compute even one iteration of Step 2 by hand. One has to use numerical approximation and maximization routines on computers.

It is known that the value function inherits monotonicity and concavity properties of the one-period return function F. In addition, Benveniste and Scheinkman (1979) showed that the value function is once differentiable under fairly general conditions. See Stokey and Lucas (1989) for more on the properties of the value function.

Dynamic programming enables researchers to analyse interesting economic problems that cannot be solved otherwise. Thus, it is no surprise that Bellman equations are widely used in economics. Below, I provide two examples of such usage.

Example 1 ***Neoclassical Growth Model***
Brock and Mirman (1972) set up a neoclassical growth model with log preference and full

depreciation. This example is one of the few cases where one can actually solve the Bellman equation by hand, using the value function iteration method. The planner's problem is to maximize $\sum_{t=0}^{\infty} \beta^t \ln(c_t)$, subject to the resource constraint of $c_t + k_{t+1} \leq Ak_t^{\alpha}$, with $A > 0$, $\alpha \in (0, 1)$ and $\beta \in (0, 1)$. In this problem, k_t is the state variable and c_t is the control, with $\Gamma(k) = \{c : 0 < c \leq Ak^{\alpha}\}$, $g(k_t, c_t) = Ak_t^{\alpha} - c_t$ and $F(k_t, c_t) = \ln(c_t)$. The Bellman equation for this problem is:

$$V(k) = \max_{0<c\leq Ak^{\alpha}} \ln(c) + \beta V(Ak^{\alpha} - c).$$

Let's solve the Bellman equation by iterating on the value function. Begin by guessing $V_0(k) = 0$. Following the procedure outlined above, we obtain:

$$V_1(k) = \ln(Ak^{\alpha}) = \ln(A) + \alpha \ln(k), V_2(k)$$
$$= \ln\frac{A}{1+\alpha\beta} + \beta\ln(A) + \alpha\beta\ln\frac{\alpha\beta A}{1+\alpha\beta} + \alpha(1+\alpha\beta)\ln(k).$$

Iterating onwards and using the summation formula for geometric series, we arrive at:

$$V(k) = \frac{1}{1-\beta}\left\{\ln(A(1-\alpha\beta)) + \frac{\alpha\beta}{1-\alpha\beta}\ln(A\alpha\beta)\right\} + \frac{\alpha}{1-\alpha\beta}\ln(k).$$

The optimal decision rule can now be easily computed: $c^* = h(k) = (1 - \alpha\beta)Ak^{\alpha}$.

Example 2 ***Consumption Smoothing***

Our discussion of Bellman equations up to this point has been limited to deterministic models. However, as long as the objective function is additively separable over time and is linear in probability, we can easily accommodate uncertainty. For example, Miller (1974) analyses a consumer's utility maximization in the face of a stochastic income stream using dynamic programming. What follows is an adapted version of Miller's model.

Think of an infinitely lived consumer or dynasty that maximizes the discounted sum of the expected utility stream. The consumer derives utility from consumption c_t, and we denote the utility function with $U(c_t)$. Her income follows a Markov process $\{y_t\}_{t=0}^{\infty}$, and the distribution of y_{t+1} given y_t is represented by the cumulative density function $G(y_{t+1}|y_t)$. We assume that $y_t \in [0, y_{\max}]$, $\forall t$. The consumer's discount factor is $\beta \in (0,1)$ and the market interest rate is r. It is assumed that $\beta(1 + r) < 1$. She can borrow and lend at the market interest rate, but her debt cannot exceed $B_{\max} < \infty$. We denote her asset holdings at the beginning of period t with a_t. To be precise, c_t and a_t are measurable functions with respect to the σ-algebra generated by the income process. For notational convenience, we suppress this history dependence. Now we write down the consumer's problem:

$$\max_{\{c_t\}_{t=0}^{\infty}} E_0 \sum_{t=0}^{\infty} \beta^t U(c_t),$$

s.t. $c_t + \frac{a_{t+1}}{1+r} \leq a_t + y_t, a_t \geq -B_{\max}$ and $y_{t+1} \sim G(y_{t+1}|y_t)$, with a_0 and y_0 given.

To obtain a recursive formulation, it must be noted that (a_t, y_t) are the relevant state variables. Without loss of generality, assume that there is no borrowing. The Bellman equation for this consumer's problem is then:

$$V(a, y) = \max_{0 \leq c \leq a+y} U(c) + \beta \int V((1+r)(a+y-c), y')dG(y'|y).$$

Unlike in the first example, this Bellman equation cannot be solved by hand in general, and necessitates numerical methods.

See Also

▶ Dynamic Programming

Bibliography

Bellman, R. 1957. *Dynamic programming*. Princeton: Princeton University Press.

Benveniste, L., and J. Scheinkman. 1979. On the differentiability of the value function in dynamic models of economics. *Econometrica* 47: 727–732.
Bertsekas, D.P. 1976. *Dynamic programming and stochastic control*. New York: Academic Press.
Blackwell, D. 1965. Discounted dynamic programming. *Annals of Mathematical Statistics* 36: 226–235.
Brock, W.A., and L. Mirman. 1972. Optimal economic growth and uncertainty: The discounted case. *Journal of Economic Theory* 4: 479–513.
Ljungqvist, L., and T.J. Sargent. 2004. *Recursive macroeconomic theory*. 2nd ed. Cambridge, MA: MIT Press.
Miller, B.L. 1974. Optimal consumption with a stochastic income stream. *Econometrica* 42: 253–266.
Stokey, N.L., and R.E. Lucas Jr. 1989. *Recursive methods in economic dynamics*. Cambridge: Harvard University Press.

Ben Porath, Yoram (1937–1992)

Reuben Gronau

Keywords

Ben Porath, Y.; Child mortality; Family economics; Fertility; Human capital investment; Population growth; Women and work

JEL Classifications

B31

Yoram Ben Porath's paper 'The Production of Human Capital and the Life Cycle of Earnings' (1967) is still regarded as one of the path-breaking papers in the economics of human resources. Following Mincer and Becker, the paper uses the framework of optimum control to analyse the joint decision of investment in human capital and market work over the life cycle. Diminishing marginal productivity in the investment process results in the process being spread over a lengthy period of time. A shrinking horizon results in the time devoted to the investment diminishing over the life cycle, an increasing fraction of time being diverted to market work. The model, part of Ben Porath's doctoral dissertation, provides an elegant economic explanation for the concentration of formal studies (that is, 'full-time' investment) early in life, and the concave shape of the age-earning profile.

Ben Porath's MA thesis (1966) was the most comprehensive economic study of the Arab labour force and the Arab sector in the Israeli economy at the time of its composition. Like his doctorate, it reflects Ben Porath's lifetime interest in the interaction between human resources and growth. In a series of studies on fertility patterns in Israel he explored the substitution between quality and quantity, sex preferences and family size (1976, 1981), the effect of child mortality on family size (1976), and the interaction between fertility and women's labour supply (1985), combining theory and empirical research.

Ben Porath's interest in the economics of fertility led him to widen the scope of investigation, focusing on the economic functions of the family. In his 1980 essay 'The F-connection: Families, Friends and Firms and the Organisation of Exchange' he explored the social and economic role of families, contrasting the exchange taking place within the family (or other small socially knit groups) which are characterised by 'specialization by identity' and the conventional view of market exchange between anonymous buyers and sellers. In a world of imperfect information the transactional advantages of trade within a small group plays an important role in explaining the shifting border between the family and the market.

In 1979, when Ben Porath became the director of the Maurice Falk Institute for Economic Research in Israel, he initiated a comprehensive study of the economy of Israel, an economy plagued by an uncontrollable inflation and halting growth. In the opening paper of the volume that he edited, *The Israeli Economy: Maturing through Crisis* (1986), he returned to tackle the question that puzzled him throughout his career – the interaction between output and population growth: is population growth the engine of output growth, or does output growth encourage immigration?

Yoram Ben Porath was born in Tel Aviv in 1937. He started his studies in economics at the Hebrew University in Jerusalem in 1957, and received his Ph.D. from Harvard in 1967, studying with Simon Kuznets. In 1986 he was elected

Deputy Provost of the Hebrew University, and later became Provost. In 1990 he was elected president of the university. In 1992, during his term as president, he was killed in a car accident.

Selected Works

1966. *The Arab Labor Force in Israel.* Jerusalem: Maurice Falk Institute for Economic Research in Israel.
1967. The production of human capital and the life cycle of earnings. *Journal of Political Economy* 75(pt. 1), 352–365.
1976a. Fertility response to child mortality: Microdata from Israel. *Journal of Political Economy* 84, S163–S178.
1976b. (With F. Welch.) Do sex differences really matter? *Quarterly Journal of Economics* 90, 285–307.
1980. The F-connection: Families, friends, and firms and the organisation of exchange. *Population and Development Review* 6, 1–30.
1981. (With F. Welch.) On sex preferences and family size. In *Research in Population Economics*. Vol. 2, ed. J. Simon. Greenwich, CT: JAI Press.
1985. (With R. Gronau.) Jewish mother goes to work: Trends in the labor force participation of women in Israel, 1955–80. *Journal of Labor Economics* 3(pt. 2), S310–S327.
1986. *The Israeli Economy: Maturing through Crises.* Cambridge, MA: Harvard University Press.

Bentham, Jeremy (1748–1832)

Philip Schofield

Abstract

Jeremy Bentham, English philosopher and reformer, was the founder of classical utilitarianism, the doctrine that an action was morally right to the extent that it promoted the greatest happiness of the greatest number. In Bentham's hands, the principle of utility provided a critical standard by which to test the value of existing practices, laws, and institutions, and to suggest reform and improvement. His basic premise in political economy was that wealth would be most effectively produced where the individual was left free from government intervention, though government had a crucial role in providing the background conditions of security without which civilized life was impossible.

Keywords

Abundance; Bentham, J.; Civil law; Colonies; Consumption taxation; Democracy; Diminishing marginal utility; Economic freedom; Equality; Escheat; Ethics; Expectations; Free press; Happiness; Inflation; Moneylenders; National debt; Panopticon; Political economy; Population growth; Possession; Principle of utility; Property; Property taxation; Psychological hedonism; Public debt; Punishment; Redistribution of income and wealth; Security; Sinister interests; Sinking fund; Subsistence; Usury; Utilitarianism; Wealth; Well-being

JEL Classifications

B31

Jeremy Bentham, English philosopher and reformer, was the founder of classical utilitarianism, and, thereby, arguably the founder of the modern discipline of economics.

Bentham was born in Church Lane, Houndsditch, London on 15 February 1748. His father Jeremiah Bentham (1712–1792) was a solicitor, with a practice in the Court of Chancery, and wealthy and important clients in the City of London. Of his six siblings, only one younger brother Samuel (1757–1831) survived into adulthood, becoming a prominent naval architect and engineer. His mother Alicia died on 6 January 1759. A precocious child, he was educated at Westminster School until 1760 when his father entered him, at the age of 12, into the University

of Oxford, where he graduated in 1764, reputedly the youngest person ever to have done so. In the meantime, in accordance with his father's wish to see him pursue a career in the law, he had entered Lincoln's Inn in 1763, and was admitted to the bar in 1769. In that same year, however, he convinced himself that he should not practise law but rather devote himself to legal reform. Bentham thought of himself as 'the Newton of legislation' – just as Isaac Newton (1642–1727) had brought order to the physical sciences, so would Bentham to the moral sciences. He adopted the principle of utility (an action was judged to be morally right to the extent that that it promoted the greatest happiness of the greatest number) as a critical standard by which to test the value of existing practices, laws, and institutions, and to suggest reform and improvement. He set about composing a comprehensive code of laws, to which his best-known work, *An Introduction to the Principles of Morals and Legislation* (printed 1780, published 1789), was intended to form a preface. He announced that his enterprise was 'to rear the fabric of felicity by the hands of reason and of law' (Bentham 1970, p. 11).

Principle of Utility

Bentham's critical standard, the principle of utility, was based on the psychological insight that sentient creatures were motivated by a desire for pleasure and an aversion to pain. An individual had a motive to perform an action – or, put another way, had an interest in performing it – if he expected to gain some pleasure or avert some pain from doing so, and the greater or more valuable the pleasure experienced or pain averted, the stronger the motive or greater the interest. The value of a pleasure or pain was determined by its quantity, which, in the case of a single individual was a product of its intensity, duration, certainty, and propinquity. Where the value of a pleasure or pain was considered in relation to more than one person, then, in addition to these circumstances, the circumstance of extent, that is, the number of persons affected by it, had to be taken into account. At this point, a statement of psychological fact became a statement of moral science. An act was morally good if, after calculating all the pains or pleasures produced in the instance of every individual affected, the balance was on the side of pleasure, and morally evil if on the side of pain. Psychology and ethics were both founded on, and therefore linked by their relation to, pleasure and pain. Hence, Bentham's statement that, 'Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do.' The 'sovereign masters' of pain and pleasure not only accounted for human motivation, 'govern[ing] us in all we do, in all we say, in all we think', but also provided 'the standard of right and wrong'. (Bentham 1970, p. 11).

Panopticon

The middle part of Bentham's life, from about 1790 to 1803, was dominated by his attempt to build a panopticon prison in London. The panopticon design was the brainchild of Bentham's brother Samuel, when employed in the 1780s on the estates of Prince Grigoriy Aleksandrovich Potemkin (1724–1791) at Krichev, in Russia. He found that, by organizing his workforce in a circular building, with himself at the centre, he could supervise its activities more effectively. On a visit to his brother in the late 1780s and seeing the design, Bentham immediately appreciated its potential. Enshrining the principle of inspection, the panopticon might be adapted as a mental asylum, hospital, school, poor house, factory, and, of course, prison. The prison building would be circular, with the cells, occupying several storeys one above the other, placed around the circumference. At the centre of the building would be the inspector's lodge, which would be so constructed that the inspector would always be capable of seeing into the cells, while the prisoners would be unable to see whether they were being watched. The activities of the prisoners would be transparent to the inspector; his actions, in so far as the prisoners were concerned, were hidden behind a veil of secrecy. On the other

hand, it was a cardinal feature of the design that the activities of the inspector and his officials should be laid open to the general scrutiny of the public, who would be encouraged to visit the prison. When the panopticon scheme effectively collapsed in 1803, Bentham was left embittered by what he regarded as the bad faith of successive ministries, and he became increasingly committed to political radicalism.

Defence of Usury

While in Russia, Bentham composed *Defence of Usury* (1787), which proved to be one of his most successful attempts to influence economic policy. Bentham greatly admired Adam Smith's *Wealth of Nations*, which he studied in detail. He was not, however, an uncritical admirer, and argued that Smith had contradicted his own free market principles by defending the legal prohibition against exorbitant rates of interest. Countering the popular sentiment which condemned the moneylender for his avarice and pitied the borrower, Bentham argued that the former embodied the virtues of frugality, thrift, and prudence, and the latter, whether described as an entrepreneur or a prodigal, should be allowed to decide for himself whether to enter into a particular money bargain. In other words, Bentham saw no reason why the freedom of commerce should not be extended to the lending and borrowing of money. At the same time, Bentham defended the projector from the criticisms of Smith, who had linked the projector with the prodigal, and contrasted both with the sober person. The projector (and Bentham, with his panopticon prison scheme, placed himself in this category) promoted utility by improving existing products and processes or by inventing new and better ones: in short, projectors were the agents of progress.

Political Economy and the Four Sub-ends of Utility

Bentham's most intense period of work on questions of political economy took place between 1793 and 1801. Political economy, like all other fields of knowledge, had a place in Bentham's classification of knowledge, and consequently a place in his conception of a comprehensive code of laws. It was the task of the utilitarian legislator to introduce measures which would increase the overall happiness (understood in terms of a balance of pleasure over pain), or, more centrally, which would prevent a decrease in happiness. This task would be undertaken by promoting what Bentham termed the four sub-ends of utility – subsistence, abundance, security and equality – using, where appropriate, sanctions (punishments and rewards), themselves composed of pain and pleasure, to discourage actions detrimental to the happiness of the community, and (to a lesser extent) to encourage those which were beneficial. More specifically, it was the task of the civil law to distribute rights and duties in such a way as to promote the four sub-ends of utility. Security consisted in the protection of the basic interests of the individual – his person, property, reputation, and condition in life – which constituted a major component of his well-being. Security was closely related to the notion of expectations, for it involved both the present possession and the future expectation of possessing the property or other subject-matter in question. Without security, and thus the confidence to project oneself and one's plans into the future, there could be no civilized life. In short, security was a product of law, resulting from the imposition of rules on conduct.

The subject of political economy was more particularly concerned with subsistence and abundance, though the significance of security and equality should not be overlooked. For instance, without the security provided by law, no one would have an incentive to labour, and, therefore, to create wealth (abundance). Moreover, abundance itself was a security for subsistence, that is, the minimum quantity of resources which an individual needed to survive. Indeed, it was subsistence which had a prior claim on all resources in that an individual could be happy only if he were alive. Once wealth had been created, the principle of equality – in essence, the principle of diminishing marginal utility – demanded that it be distributed equally. Bentham argued that, if subsistence required £10 per annum, the most

important £10 which an individual could possess was the first £10. Thereafter, each increment of £10 was worth something less than the previous increment. To put this another way, £10 given to an individual who had nothing constituted the difference between life and death, whereas £10 given to a rich man made hardly any difference at all. Bentham did not, however, advocate the levelling of property, for two reasons. First, if everyone began one morning with the same amount of property, by the end of the afternoon the intervening transactions would see inequality re-established. Second, the levelling of property would constitute an attack on security. Indeed, security, with its attendant expectations, was so important, that it was only in exceptional circumstances, such as providing subsistence to those who might otherwise starve to death, that it was legitimate to redistribute resources, and even here Bentham partly justified the redistribution on the grounds of security, in that such redistribution would render the property of the rich less liable to violent invasion by the poor.

In relation to abundance, or the creation of wealth, Bentham's basic principle was that of economic freedom. Each individual was most likely to be the best judge of his own interest, since he was most likely to be best informed about his own peculiar circumstances, and most likely to be motivated to act on that information in order to maximize his wealth, and thence his happiness. In a large number of areas in which government had traditionally intervened in economic matters, its intervention was counterproductive. Trade bounties, prohibitions, monopolies, and encouragements to population growth belonged to what Bentham termed the 'non-agenda' (although there might always be exceptions). Taking his lead from Smith, Bentham argued that since trade was limited by capital, government could not favour one branch of trade unless it discouraged another branch, since the capital applied to the former must be taken from the latter. In general, government was best advised not to interfere with the economy, and this included interference in the form of taxation. The imposition of taxation was a form of coercion, and all coercion was an evil in itself. As Bentham remarked: 'The best use that government can make of money in the hands of the lawful possessors is: to leave it where it is' (Bentham 1989, p. 251). He argued that, in order to judge the utility of any element of public expenditure, one needed to compare the benefits produced by the expenditure with the burden produced by imposing an equivalent degree of taxation in the most aggravated form in which taxation was imposed. Hence, he recommended the immediate repeal of several particularly burdensome taxes – for instance those on legal proceedings, medicines, insurance, and newspapers (the latter constituting a tax on information). The taxation which remained should be imposed where there existed an ability to pay. Hence, the best form of taxation was that on consumption, followed by that on property and the transfer of property. As an alternative source of public revenue, he advocated a revival of the medieval practice of escheat, whereby the state appropriated property where there was no other than a collateral heir. The money raised would be earmarked for a sinking fund, which would eventually redeem the national debt. The appropriation of collateral successions was a measure which Bentham believed could reconcile the otherwise conflicting demands of security and equality. Providing that individuals knew in advance that their potential to inherit would be limited according to law, they would not suffer any disappointed expectations, and their security would not be infringed. Apart from providing the background conditions of security which ensured that economic actors had the incentives to accumulate wealth (for instance security of person and property), there was, nonetheless, a limited 'agenda' for government, for instance to establish corn magazines to provide a security against dearth, to provide information, and to commission and disseminate research.

Monetary Regulation

Following the suspension of payments in *specie* at the Bank of England in 1797, Bentham turned his attention to monetary regulation, devising his annuity note scheme, with the aim of redeeming

the national debt. The annuity notes would in effect serve as paper currency, but at the same time earn compound interest, and, therefore, act as an investment. Depending on the prevailing rates of interest, holders of the notes would either use them as currency or horde them as savings. The government would issue the notes in order to buy up existing public debt, and thereafter successively reduce the rate of interest payable. The annuity notes as a circulating medium would replace an equivalent amount of bank notes, and lead to an earlier redemption of the national debt than would otherwise have been possible. It seems that Bentham abandoned the scheme because he did not, to his own satisfaction, solve the problem of inflation, which, he feared, would stifle the growth of national wealth and unfairly reduce the real value of fixed incomes.

In 1801 Bentham calculated that prices had increased by 50 per cent since 1760. He argued that this inflation had been caused by an increase in the amount of paper money in circulation. This increase was to be welcomed in that it represented a growth in national prosperity. However, it also represented an unfair tax on fixed incomes, and threatened a general bankruptcy. His remedy was to limit and to tax the issue of paper money by provincial banks, who were prone to over-issue bank notes since this was the main source of their profit. In return, a licensing system would be introduced which would, in effect, grant a monopoly to existing banks. In December 1801, in the extraordinary circumstances brought about by scarcity and dearth of provisions, he came to advocate legislative intervention in the economy in the form of the statutory imposition of a maximum price for wheat. This would have the immediate effect of bringing relief to the poor and security to the propertied, in that it would avoid the creation of a potentially revolutionary situation fuelled by the discontent of the destitute. Scarcity, he argued, could only permanently be remedied by the establishment of corn magazines and the promotion of emigration, both of population and of capital. In short, while favouring economic liberty as a leading principle, he was always prepared to consider state intervention should the principle of utility demand it.

Colonies

Bentham's opposition to the holding of colonies was grounded initially on economic arguments, though he later developed political and constitutional objections to the practice. Given that the trade of a nation was limited by the quantity of capital it possessed, he argued that colony-holding could not bring any economic advantages. The extension of markets which the acquisition of colonies appeared to provide did not in itself affect the amount of trade. New markets were advantageous only to the extent that the profit made upon the capital employed in the new trade was greater than the profit made on the established trade. It was unlikely that the distant markets represented by colonies would offer a higher rate of return than those closer to home. Any benefit from a trade monopoly imposed on the produce of the colony was illusory, since a monopoly could not force the price of a commodity lower than the level to which it would be driven by competition, and it could not force anyone to produce a commodity at a loss. Finally, to the argument that trade with colonies was a source of revenue, Bentham responded that revenue could be raised on goods exchanged with all other countries, not just colonies, providing of course that the duties were not so high as to make smuggling attractive. The emancipation of colonies would also save the mother country the massive expense of defending them, particularly in time of war. Nonetheless, there were certain circumstances in which Bentham was prepared to defend the establishment of colonies. He approved the colonization of vacant lands in response to the pressure of population growth and the existence of an excess of capital in the mother country, and of colonial rule in countries where the native rulers were unfit to govern. The benefits, however, accrued to the colonists, and not to the mother country, and he recommended that dominion should be relinquished as soon as was practicable.

Political Reform

By the 1820s Bentham was convinced that the only regime with an interest in enacting good legislation

was a representative democracy. A crucial development took place around 1804 with the emergence in Bentham's thought of the notion of sinister interests, that is, the systematic development of the insight that rulers wished to promote not the happiness of the community, but their own happiness. There was no point in showing rulers what the best course of legislation might be unless they had an interest in adopting it. Only a legislature elected by a democratic suffrage had such an interest. Following the quashing of the panopticon scheme in 1803, Bentham became convinced that nothing worthwhile could be achieved through the existing political structure in Britain, or through similar regimes elsewhere. Having concentrated on questions of law reform from 1803, he was in the summer of 1809 prompted to compose material on political reform, eventually bearing fruit in *Plan of Parliamentary Reform* (1817). In this work he called for universal manhood suffrage (subject to a literacy test), annual parliaments, equal electoral districts, payment of MPs, and the secret ballot. Bentham then went a stage further and drew up a blueprint for representative democracy which would have abolished the monarchy, the House of Lords and any other second chamber, and all artificial titles of honour, and would have rendered government entirely open and, he hoped, fully accountable. These proposals were developed in astonishing detail in the magisterial *Constitutional Code* (partly printed 1827 and 1830, partly published 1830).

For Bentham the key principle of constitutional design was to ensure the dependence of rulers on subjects. Instead of the traditional theory of the separation of powers, he proposed lines of subordination, based on the ability of the superior to appoint and dismiss (in Bentham's terminology to locate and dislocate) the inferior, and to subject the inferior to punishment and other forms of 'vexation'. The supreme power or sovereignty in the state would be vested in the people, who held the constitutive power. Immediately subordinate to the people would be the legislature, elected by universal manhood suffrage, and subordinate to the legislature would be the administrative (that is, the executive) and judicial powers. The system of representative democracy was not an end in itself – the end was the greatest happiness – but was an indispensable means to that end, in that it was only under such a constitution that effective measures could be implemented to secure the good behaviour (appropriate aptitude) of officials and minimize the expense of government. The securities for official aptitude – otherwise termed securities against misrule – included the exclusion of factitious dignities (titles of honour), the economical auction (whereby officials made bids for the salary attached to the office), subjection to punishment at the hands of the legal tribunals of the state, the requirement to pass an examination, and, most importantly, publicity. Bentham went to great lengths to ensure that government would be open to public scrutiny, and thence subject to the force of the moral or popular sanction operating through the public opinion tribunal, which consisted in all those who commented on political matters, and of whom newspaper editors were the most important. Bentham saw the freedom of the press as a vital bulwark against misrule: hence his proposal to encourage the diffusion of literacy by making the suffrage dependent on a literacy test. These measures were intended to ensure that rulers would be so situated that the only way they could promote their own interest was by promoting the interest of the community.

Death and Afterwards

Having lived in Lincoln's Inn from 1769 to 1792, he had then inherited his father's home in Queen's Square Place, Westminster, where he died on 6 June 1832. It was Bentham's wish that his body be dissected for the advancement of medical science, and that his remains then be used to create an 'auto-icon' or self-image. Bentham's auto-icon, assembled by his surgeon Thomas Southwood Smith (1788–1861), and consisting in a waxwork head mounted on Bentham's articulated skeleton and wearing his clothes, is now kept at University College London.

See Also

► Utilitarianism and Economic Theory

Selected Works

The Bentham Project, University College London, is preparing a new authoritative edition of *The Collected Works of Jeremy Bentham*, which, it is estimated, will run to 68 volumes. The 26th appeared in February 2006. The following volumes have been most extensively drawn upon in the compilation of this article:

1970. *An introduction to the principles of morals and legislation*. Edited by J.H. Burns and H.L.A. Hart. London: Athlone Press.
1977. *A comment on the commentaries and a fragment on government*. Edited by J.H. Burns and H.L.A. Hart. London: Athlone Press.
1989. *First principles preparatory to constitutional code*. Edited by P. Schofield. Oxford: Clarendon Press.
1998. *'Legislator of the world': Writings on codification, law, and education*. Edited by P. Schofield and J. Harris. Oxford: Clarendon Press.

Where cited works have not appeared in *The Collected Works*, the standard source is the so-called Bowring edition: *The Works of Jeremy Bentham, published under the superintendence of his executor, John Bowring*, 11 vols. Edinburgh: William Tait, 1843. The standard source for Bentham's economic thought is *Jeremy Bentham's Economic Writings*, 3 vols., ed. W. Stark. London: George Allen & Unwin, 1952–54. A new authoritative edition is greatly needed.

Bibliography

Dinwiddy, J. 2004. *Bentham: Selected writings of John Dinwiddy*. Edited by W. Twining. Stanford: Stanford University Press.
Harrison, R. 1983. *Bentham*. London: Routledge and Kegan Paul.
Hart, H.L.A. 1982. *Essays on Bentham: Jurisprudence and political theory*. Oxford: Clarendon Press.
Kelly, P.J. 1990. *Utilitarianism and distributive justice: Jeremy Bentham and the civil law*. Oxford: Clarendon Press.
Lieberman, D. 2000. Economy and polity in Bentham's science of legislation. In *Economy, polity, and society: British intellectual history, 1750–1950*, ed. S. Collini, R. Whatmore, and B. Young. Cambridge: Cambridge University Press.
Postema, G.J. 1986. *Bentham and the common law tradition*. Oxford: Clarendon Press.
Rosen, F. 2003. *Classical utilitarianism from Hume to Mill*. London: Routledge.
Schofield, P. 2006. *Utility and democracy: The political thought of Jeremy Bentham*. Oxford: Oxford University Press.
Semple, J. 1993. *Bentham's prison: A study of the panopticon penitentiary*. Oxford: Clarendon Press.
Warke, T. 2000. Multi-dimensional utility and the index number problem: Jeremy Bentham, J.S. Mill, and qualitative hedonism. *Utilitas* 12: 176–203.

Bequests and the Life Cycle Model

John Laitner

Abstract

The standard life cycle model emphasizes a household's concerns over events within its lifetime, including providing for its own retirement and for its young children. However, in a more elaborate formulation, the household may care about its descendants when they are grown just as when they are young, causing the household to want to leave bequests. Its time horizon may expand to a dynastic scale, and new public policy implications, including so-called Ricardian neutrality, may emerge. Alternatively, bequests may signal non-market exchanges between parents and their adult children, perhaps arising to mitigate transactions costs or informational asymmetries.

Keywords

Altruistic bequests; Annuities; Assortative mating; Bequests; Bequests and the life cycle model; Implicit contracts; Infinite horizons; *Inter vivos* transfers; Joy of giving; Life cycle hypothesis; Life-cycle model; Non-market exchange; Representative agent; Retirement; Ricardian neutrality; Strategic behaviour

JEL Classifications

D4; D10

In the life-cycle model of household behaviour, each household expects a lifetime pattern of rising earnings in youth and middle age followed by retirement. Hence, households plan to save in their first segments of life in order to build resources to dissave, and from which to accrue interest income, during the last (Modigliani 1986). The framework easily incorporates children, with consumption early in a household's life driven higher and saving for retirement perhaps delayed until middle age (Tobin 1967). In a standard life-cycle model, parents plan for their own life and assume financial responsibility for their children until the latter reach adulthood (say, age 18 or 22) – but not beyond. Elaborations of the framework, on the other hand, extend parental concern, or interest in non-market transactions, to encompass a household's grown children. Such elaborations expand the scope of the life-cycle model to include bequests.

Conceptually, there are at least three broad categories of models in which bequests play a role. The first, which is often called the 'altruistic model', assumes that parents care about the well-being of their grown children. The second, which one might call the 'joy of giving model', assumes that parents derive pleasure from making transfers to their adult children's households but that the pleasure is not specifically dependent upon the children's utility gain. In the third formulation, parent-to-child emotional and social ties favour and facilitate non-market exchanges that may generate bequests – for example, bequests may emerge as payments to heirs for personal services rendered.

Altruistic Model

A model with 'altruistic bequests' (Becker 1974; Barro 1974) extends to grown children parental concerns for minor children typical of standard life-cycle analyses.

Consider a specific example in which each household has one adult, raises one child, and lives two periods. Suppose that a household begun at time t has earnings y_t in youth but is retired in old age. It rears its child during its first stage of life; the child initiates its own household thereafter, with the descendant household passing its first stage of life as the parent household lives through its second stage. The time-t parent chooses consumption c_t^1 and c_t^2, respectively, for its two stages of life; derives utility $u\left(c_t^1, c_t^2\right)$ from this consumption; inherits i_t in youth; and transfers i_{t+1} in old age to its adult child. Let the interest rate be r. Given i_t and i_{t+1}, the parent household's *lifetime utility* is $U(.)$ such that

$$U(i_t + i_{t+1}, y_t) \equiv \max_{c_t^1, c_t^2} u\left(c_t^1, c_t^2\right)$$

$$\text{subject to}: \; c_t^1 + \frac{c_t^2}{1+r} + \frac{i_t + 1}{1+r} \leq i_t + y_t$$

Let the parent household care δ times as much about its adult child's lifetime utility as about its own, δ^2 times as much about its grandchild's lifetime utility, and so on. Then the parent household's *dynastic utility* is

$$\sum_{s=0}^{\infty} \delta^s \cdot U\left(i_{t+s}, i_{t+s+1}, y_{t+s}\right)$$

If $y_t = y$ all t, if institutions force bequests to be nonnegative, and if descendant households share the same preference ordering, we can characterize the time-t parent household's dynastic utility as $V(i_t, y)$ with

$$V(i_t, y) = \max_{i_t+1 \geq 0} \{U(i_t, i_{t+1}, y) + \delta \cdot V(i_{t+1}, y)\}. \tag{1}$$

If $\delta = 0$, we have a 'pure life–cycle model'; if $\delta > 0$, we have an altruistic model in which positive bequests may emerge.

Laitner (1992) studies a second altruistic formulation, one allowing heterogeneous earning abilities. In terms of the framework above, a parent household with earnings y_t may know the random variable, say, $\tilde{y}$, from which the earnings of its descendants will be (independently, in the simplest case) sampled, but the parent cannot observe the sampling outcomes as it makes its bequest plans. Then dynastic utility is

$$V(i_t, y) = \max_{i_{t+1} \geq 0} \{U(i_t, i_{t+1}, y) + \delta \cdot E[V(i_{t+1}, \tilde{y})]\}, \quad (2)$$

where $E[.]$ is the expectations operator.

Conceptually, a model with altruistic bequests provides an extension of the life-cycle model's parental concern for minor children's well-being to a more or less symmetric concern for grown children. Empirically, bequests and *inter vivos* transfers to adult children certainly occur in practice (Modigliani 1986; Kotlikoff 1988). The formulation with heterogeneous earnings predicts that bequests need not be universal but are most likely in the case of very prosperous parents. Social commentators frequently criticize bequests as a source of inequality, and the second point in the preceding sentence shows how bequests can contribute to cross-sectional dispersion of private wealth holdings. Bequests may have played a larger role in national wealth accumulation in the past, when long retirement spells were perhaps less common (Darby 1979), and a model with both life-cycle saving and altruistic bequests can provide a framework for analysing the change (Laitner 2001).

Loans for education fail to generate collateral for creditors; hence, parental and/or public support may be important for ensuring efficient educational investment. Since benefits of education last long into adulthood, the model with altruistic bequests provides a logical framework for studying parental contributions (for example, Tomes 1981). For instance, suppose that a child's earnings are an increasing, concave function $f(.)$ of ability, a, and parental support for education, e, in the child's youth: $y_{t+1} = f(a_t, e_t)$. With homogeneous agents, $a_t = a$ all t, and (1) becomes

$$V(i_t, y) = \max_{i_{t+1} \geq 0, e_t \geq 0} \{U(i_t, i_{t+1} + e_t \cdot (1 + r), y) + \delta \cdot V(i_{t+1}, f(a, e_t))\}. \quad (3)$$

Then $i_{t+1} > 0$ ensures efficient provision of education e_t regardless of the degree of parental concern for the child, δ. If, on the other hand, the tangible bequest is zero, investment in education can be inefficiently low.

A second prominent application of the altruistic model relates to fiscal policy. In a standard life-cycle model, when government turns from tax to deficit finance, national consumption may rise for a time, and the economy's long-run capital intensity may decline. Reformulating the life-cycle model to include altruistic bequests can overturn this result (for example, Barro 1974). Debt service and repayment for current government borrowing may extend far beyond the life span of existing households, but not beyond the time horizon of dynasties. Maximization in (1) may yield an outcome in which the non-negativity constraint never binds, and Barro (1974) shows that in that case tax and deficit finance may have identical implications for aggregate consumption, capital accumulation, and interest rates. The latter equivalence is often referred to as 'Ricardian neutrality'. (With heterogeneity of agents, as in formulation (2), non-negativity constraints will, on the other hand, tend to bind for some households – Laitner 1992 – and then outcomes resembling Ricardian neutrality, while still possible, may be more in doubt – for example, Bernheim 1987.)

Recent dynamic general equilibrium analyses of long-run growth and business cycles frequently employ the so-called 'representative agent' paradigm. Utility maximization over an infinite time horizon for a set of identical agents determines desired private consumption, saving, and labour supply. It seems fair to say that the life-cycle model with altruistic bequests, as in Barro (1974) and related papers, provides the most basic motivation for this approach.

Turning to empirical findings, the widespread existence of bequests (and *inter vivos* gifts) within family lines is well established (Modigliani 1986; Kotlikoff 1988). The pure life-cycle model does not seem able to explain as much national wealth as we see, and estate building seems a plausible explanation for the remainder (Kotlikoff 1988). However, despite some consistency with the altruistic model, empirical evidence often seems to fail to support the implications of pervasive Ricardian neutrality (for example, Altonji et al. 1992, 1997). Long-standing evidence that households with multiple children tend in practice to divide their bequests equally (for example, Menchik 1988)

also seems contrary to implications of the simplest versions of the altruistic model. Perhaps altruistic bequest behaviour is, in practice, concentrated among the highest-income households (as might be implied by formulation (2)).

Conceptually, as one considers couples instead of single parents, dynasties will interact through marriage. Assortative mating can preserve the logic of the analysis of the parthenogenetic theoretical construct (Laitner 1991). Mating patterns that are random theoretically could, in contrast, expand to an overwhelming degree the scope of interpersonal connections that 'neutralize' incentives for self-interested behaviour (Bernheim and Bagwell 1988).

The preceding formulations assume that a parent cares about his child but that the reverse is not true. A number of papers analyse two-sided altruism. Implicitly, in fact, all formulations with altruistic transfers are two sided – in model (1), for example, the parent cares about his child's utility relative to his own with a ratio of weights δ:1, while the child cares about his parent's utility relative to his own with weights in a ratio of 0:1. Unless parents and children agree on each other's relative importance, strategic behaviour may arise if agents have sufficient latitude in their set of feasible actions. In Laitner (1988), for instance, though parents and children care about each other, each may care less about the other than about itself – in which case a parent with low earnings may intentionally limit his life-cycle saving in youth in order to induce a larger transfer from his child during his retirement.

In the simplest life cycle model, a household saves before retirement in order to preserve an even level of consumption for the remainder of its life. An altruistic model extends the time frame of such behaviour: a household may use bequests (and *inter vivos* gifts) to promote evenness of consumption for its entire family line.

Joy of Giving Model

A joy-of-giving model provides a donor with pleasure that is independent of recipient utility and outside resources. For example, our two-period household above might solve

$$\max_{i_{t+1}\geq 0} \{U(i_t, i_{t+1}, y_t) + W(i_{t+1})\}, \qquad (4)$$

with the new function $W(.)$ being unrelated to lifetime utility $U(.)$ or to recipient earnings y_{t+1}. In this approach, the parent household has preferences over its own lifetime consumption and the size of the bequest that it provides to its offspring, rather than over the descendant's consumption or utility. An example is Blinder (1974).

A possible advantage of this framework is that it does not require as great an ability on the part of donors to manifest empathy and rationality as the altruistic model. Another advantage is its analytic simplicity. In applications, authors may seek to specify the utility function $W(.)$ in a manner that can mimic, at least to some degree, the model with altruistic bequests (for example, Modigliani 1986).

Exchange

The emotional ties of parents and their children may lead parents to prefer attentions from their grown children over services purchased in markets. Similarly, emotional bonds, tradition, or social norms may give trades between relatives lower transaction costs than those based on market contracts. Relatives may also have more complete information about one another than anonymous market participants do. Such factors may lead parents to make transaction and insurance arrangements with their grown children, and parental payments may take the form of bequests or *inter vivos* gifts.

In traditional societies, a household's eldest son might labour on his parents' farm, supporting his parents in their old age. In return, the son might expect to inherit the farm at his parents' death. One can view such a bequest as a payment for services, and neither altruistic nor joy-of-giving impulses on the part of parents (or their son) need be determinants of the transfer's size.

Bernheim et al. (1985) provide a model in which elderly parents desire attention from their adult children, and the parents can be thought of as paying for the services through their bequest.

Many economists note the relative infrequency with which households purchase annuities. Transactions costs and adverse selection, due to private information about one's likely longevity, may be the underlying reason. In practice, parents may circumvent annuity markets by making implicit contracts with their grown children: in return for care and support in old age, the parents agree to bequeath their assets to their children. The children take the place of an insurance company: if their parents die young, the children's efforts receive generous remuneration; if the parents live a long time, their bequest may be small or non-existent, and the children's reward per hour of effort will be low. Kotlikoff and Spivak (1981) show that such arrangements can be surprisingly efficient. Friedman and Warshawsky (1990) illustrate a related point: they show that parents who have some inclination (either joy of giving or altruistic) to bequeath to their children may eschew market annuities with even modest transactions costs, preferring self-insurance, under which their children can inherit unspent parental resources.

See Also

► Inheritance and Bequests

Bibliography

Altonji, J.G., F. Hayashi, and L.J. Kotlikoff. 1992. Is the extended family altruistically linked? Direct tests using micro data. *American Economic Review* 82: 1177–1198.

Altonji, J.G., F. Hayashi, and L.J. Kotlikoff. 1997. Parental altruism and *inter vivos* transfers: Theory and evidence. *Journal of Political Economy* 105: 1121–1166.

Barro, R.J. 1974. Are government bonds net worth? *Journal of Political Economy* 82: 1095–1117.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82: 1063–1093.

Bernheim, B.D. 1987. Ricardian equivalence: An evaluation of theory and evidence. In *NBER macroeconomics annual 2*, ed. S. Fischer. Cambridge: MIT Press.

Bernheim, B.D., and K. Bagwell. 1988. Is everything neutral? *Journal of Political Economy* 96: 308–338.

Bernheim, B.D., A. Shleifer, and L.H. Summers. 1985. The strategic bequest motive. *Journal of Political Economy* 93: 1045–1076.

Blinder, A.B.. 1974. *Toward an economic theory of income distribution*. Cambridge: MIT Press.

Darby, M.R. 1979. *The effects of social security on income and the capital stock*. Washington, DC: American Enterprise Institute.

Friedman, B.M., and M.J. Warshawsky. 1990. The cost of annuities: Implications for saving behavior and bequests. *Quarterly Journal of Economics* 105: 135–154.

Kotlikoff, L.J. 1988. Intergenerational transfers and savings. *Journal of Economic Perspectives* 2: 41–58.

Kotlikoff, L.J., and A. Spivak. 1981. The family as an incomplete annuities market. *Journal of Political Economy* 89: 372–391.

Laitner, J. 1988. Bequests, gifts and social security. *Review of Economic Studies* 55: 275–299.

Laitner, J. 1991. Modeling marital connections among family lines. *Journal of Political Economy* 99: 1123–1141.

Laitner, J. 1992. Random earnings differences, lifetime liquidity constraints and altruistic intergenerational transfers. *Journal of Economic Theory* 58: 135–170.

Laitner, J. 2001. Secular changes in wealth inequality and inheritance. *Economic Journal* 111: 691–721.

Menchik, P.L. 1988. Unequal estate division: Is it altruism, reverse bequests, or simply noise? In *Modelling the Accumulation and Distribution of Wealth*, ed. D. Kessler and A. Masson. Oxford: Clarendon Press.

Modigliani, F. 1986. Life cycle, individual thrift and the wealth of nations. *American Economic Review* 76: 297–313.

Tobin, J. 1967. Life cycle saving and balanced growth. In *Ten economic studies in the tradition of irving fisher*, ed. W. Fellner. New York: Wiley.

Tomes, N. 1981. The family, inheritance and the intergenerational transmission of inequality. *Journal of Political Economy* 89: 928–958.

Bergson, Abram (1914–2003)

Michael Ellman

Keywords

Adjusted factor cost method; Arrow's Theorem; Association for Comparative Economic Studies; Bergson A.; Bergson gap; Material Product System (MPS); Russian Research

Centre (Harvard); Social welfare function; Soviet growth record

JEL Classifications
B31

Bergson was the intellectual father of US studies of the Soviet economy during the Second World War as chief of the Russian Economic subdivision of the Office of Strategic Services (OSS). After the war he played the major role in founding the US tradition of description and analysis of Soviet economic institutions, measurement of Soviet economic growth and evaluation of that growth. He had earlier made a major contribution to the development of welfare economics. His work on the Soviet economy was marked by a combination of encyclopaedic knowledge of Soviet statistics, theoretical analysis and immense industry. It had an enormous influence on the development of US studies of the Soviet economy and established itself as the dominant paradigm in that field.

Bergson's main contribution to the study of the Soviet economy concerned the measurement of Soviet economic growth. The result of the combination of the 'propaganda of success' with Soviet economic institutions and the material product system (MPS) method of calculating national income was that the data on economic growth published by the Soviet authorities were both incredible and clearly non-comparable with the data on economic growth of other countries. Bergson both developed a method which enabled internationally comparable national income statistics and growth rates to be calculated for the USSR and applied it to the USSR for 1928–55. The method was the 'adjusted factor cost' method. In essence it consisted of adjusting actual Soviet transactions prices so as to bring them into line with the prices that would have been observed if the USSR's prices had been determined in accordance with neoclassical theory. These adjusted prices were then used as weights to aggregate the physical output series of branches and sectors of the economy as known from Soviet official data into a system of national accounts (SNA)-type aggregate. This had the great advantage of producing data comparable to SNA data and hence suitable for international comparisons. At the same time, Bergson argued, this procedure enabled a 'production potential' and possibly even a welfare interpretation to be given to the resulting national income data.

The development of this method and its application to the USSR for the period 1928–55 were enormous achievements. They clearly indicated that assessment of socialist economies did not have to remain at the level of ideological confrontation but was amenable to rational discourse and scientific inquiry. Both the method and its results were controversial. The rationality of the adjusted factor cost prices, the representativeness of the physical products selected, the huge data requirements and skilled labour inputs necessary to apply the method, the relevance of neoclassical theory for interpreting Soviet economic data, and the accuracy of the picture of the Soviet economy resulting from application of the method, all came under fire. Others used different methods of generating internationally comparable data (for example, the physical indicators method, or scaling up from net material product, NMP, to GNP using data for the missing sectors).

In welfare economics Bergson is famous for his 1938 paper which defined and discussed the concept of an individualistic social welfare function. The latter enables necessary conditions for an economic optimum to be calculated without the assumption of cardinal utility. This concept was subsequently utilized and developed by Samuelson and became an integral part of the welfare economics literature. Its usefulness remains a matter of controversy. According to Samuelson's contribution to the Bergson Festschrift it was a major contribution, a 'flash of lightning' after which 'all was light' in the hitherto extraordinarily confused subject of welfare economics. A number of opinions of a less positive kind can be found in M. Dobb (1969). Bergson also wrote on socialist economics and Arrow's Impossibility Theorem.

Besides his purely academic work on the Soviet economy, Bergson, with his OSS experience, played a major role in establishing and maintaining the close links between US academic studies of the Soviet economy and the

intelligence community and other branches of the federal government. Besides being a professor of economics for many years, first at Columbia and then at Harvard, he was director of the Harvard Russian Research Center (1964–8, 1969–70), consultant to the RAND Corporation, member and subsequently chairman of the Social Science Advisory Board of the US Arms Control and Disarmament Agency, and consultant to various federal agencies. In addition, he served as president of the Association for Comparative Economic Studies and several times testified before the US Congress.

Many years after Bergson's publications, access to Soviet economic archives demonstrated the significance and accuracy of Bergson's analysis of discrepancies in Soviet labour statistics ('the Bergson gap'). It also demonstrated the usefulness of his approach for studying the Soviet national accounts during the Second World War.

Bergson made a major contribution to 20th-century economics by establishing a school of economists who transformed the study of the Soviet economy, hitherto a reserve of partisan émigré and committed writers, into a field of sober academic inquiry.

See Also

- ▶ Social Welfare Function
- ▶ Soviet Growth Record
- ▶ Welfare Economics

Selected Works

Welfare Economics

1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52: 310–34.

1966. *Essays in normative economics.* Cambridge, MA: Harvard University Press. 1982. *Welfare, planning and employment.* Cambridge, MA: MIT Press.

Soviet Economic Institutions

1944. *The structure of Soviet Wages.* Cambridge, MA: Harvard University Press.

1964. *The Economics of Soviet planning.* New Haven: Yale University Press.

1984. Income inequality under Soviet socialism. *Journal of Economic Literature* 22: 1052–99.

Measurement of Soviet Economic Growth

1950. Ruble prices and the valuation problem. *Quarterly Journal of Economics* 64: 408–41.

1953. *Soviet national income and product in 1937.* New York: Columbia University Press.

1954. (With H. Heymann, Jr.) *Soviet national income and product 1940–48.* New York: Columbia University Press.

1961. *The real national income of Soviet Russia since 1928*. Cambridge, MA: Harvard University Press.

Evaluation of Soviet Economic Growth

1968. *Planning and productivity under Soviet socialism*. New York: Columbia University Press.

1978. *Productivity and the social system – The USSR and the west*. Cambridge, MA: Harvard University Press.

Bibliography

Birman, I. 1985. The Soviet economy: Alternative views. *Survey* 29 (2): 102–115.

Birman, I. 1986. The Soviet economy: Alternative views. *Russia* 12: 60–74.

Birman, I. 1989. *Personal consumption in the USSR and the USA*. Basingstoke: Macmillan.

Comparative Economic Studies. 2005. Performance and efficiency under socialism: Studies in honor of Abram Bergson. *Comparative Economic Studies* 47: 239–502.

Dobb, M. 1969. *Welfare economics and the economics of socialism*. Cambridge, MA: Cambridge University Press.

Hanson, P. 1971. East–West comparisons and comparative economic systems. *Soviet Studies* 22: 327–343.

Hardt, J. 2004. Abram Bergson and the development of Soviet economic studies. *Problems of Post-Communism* 51 (4): 34–39.

Harrison, M. 1996. Measuring Soviet GNP. In *Accounting for war*, ed. Mark Harrison. Cambridge, MA: Cambridge University Press.

Holzman, F. 1957. The adjusted factor cost standard of measuring national income: comment. *Soviet Studies* 9 (1): 32–36.

Khanin, G. 1993. *Sovetskii ekonomicheskii rost: analiz zapadnykh otsenok*. Novosibirsk: EKOR.

Marer, P. 1985. *Dollar GNPs of the USSR and Eastern Europe*. Baltimore: Johns Hopkins University Press.
Rosefielde, S., ed. 1981. *Economic welfare and the economics of Soviet socialism*. Cambridge, MA: Cambridge University Press. (The Bergson Festschrift – A full bibliography of Bergson's work in 1936–80 can be found on pp. 334–7 of this book.)
Samuelson, P. 2005. Abram Bergson, economist. *Economic Journal* 115 (501): F130–F133.
Wiles, P. 1955. Are adjusted roubles rational? *Soviet Studies* 7 (2): 143–160.

Berkeley, George (1685–1753)

S. Rashid

George Berkeley was an Anglican clergymen of Anglo-Irish origins who rose to be Bishop of Cloyne. He is known today principally as the philosopher of immaterialism. It is possible to look upon the economic works of George Berkeley in two different ways. First, one may consider him solely as an economic thinker and evaluate the nature and content of the ideas espoused in Berkeley's principal economic pamphlet, The Querist (1735–7), some of whose ideas are foreshadowed in the Essay towards preventing the Ruin of Great Britain (1721) and in Alciphron (1732). Secondly, one may look upon the Querist as part of the programme of economic development espoused by a number of prominent Anglo-Irishmen, a substantial number of whom were Anglican clergymen and of whom Berkeley himself was one. Viewed primarily as an economist, the two most prominent features of Berkeley's thought are his emphasis upon industry as the true source of wealth and upon the stimulation of wants as the most effective way of eliciting increased industry (Queries 1, 4, 19–21 and passim). This balanced view, partially anticipated by John Law, synthesised both the typical Mercantilist emphasis upon work as well as the stress put upon demand by such economists as Bernard Mandeville. Berkeley goes on to emphasize that economic growth would be most stimulated if the Irish would develop a taste for Irish goods (144–6). However, since such a result could not be depended upon, Berkeley was prepared to have the state intervene in order to limit the influence of fashion upon consumer tastes (13–16). Berkeley was aware that everyone may not respond to his call for increased industry and he was even willing to force such people to work (380–87). In the first edition of the Querist, Berkeley emphasized the role of the monetary system as an important catalyst for economic growth and urged the need for a National Bank in Ireland. Due to a lack of popular interest, this section was largely omitted in subsequent editions. Most of the above ideas are very much a staple of British Mercantilist writing. Berkeley does however break new ground with his philosophical analysis of the sources of wealth and by his disdain for gold and silver per se; 'Whether there ever was, is, or will be, an industrious nation poor, or an idle rich?' (Query 1), 'Whether there be any virtue in gold or silver, other than as they set people at work, or create industry?' (Query 30), as well as by his emphasis upon the welfare of the common man as the true end of economic policy; 'Whether a people can be called poor, where the common sort are well fed, clothed and lodged' (Query 2).

In a wider sense, Berkeley is to be seen as a member of a group of public-spirited Irishmen, such as Thomas Prior and the Rev. Samuel Madden, who were moved by Ireland's poverty to form a group that would help ameliorate Ireland's misery – the Dublin Society. Instead of confronting hostile English colonial policy, this group took the view that one should do whatever was feasible within the constraints set by the English. With its emphasis upon simple, practicable measures, the philosophy of the Dublin Society was very congenial to Berkeley's general aim of returning philosophy from the elite to the common man. In terms of method, Berkeley followed an iconoclastic approach, believing that a clear statement of the problems would enable common sense to perceive proper solutions. In this sense, Berkeley may be considered an anti-deductive rather than an inductive economist.

The *Querist* was very influential. Ten editions were printed even in Berkeley's lifetime. Adam Smith owned a copy and may have learned from

it. While the *Querist* continued to be read by many, such as Robert Southey and S.T. Coleridge, it was not written in a form which would endear itself to the systematizing tendencies of the classical economists. Isaac Butt tried hard to revive a Berkeleian approach in Ireland in the 1940s but failed. Nonetheless, Berkeley's genuine love for Ireland and for all the Irish people has endeared him to many, especially Irish patriots.

Selected Works

1948–57. In *The works of George Berkeley,* ed. A.A. Luce, T.E. Jessop. London: Nelson.

References

Ardley, G. 1968. *Berkeley's renovation of philosophy.* The Hague: M. Nijhoff.

Berle, Adolf Augustus, Jr. (1895–1971)

John Kenneth Galbraith

Keywords

Berle, A. A.; Corporations; Entrepreneurship; Keynes, J. M.; McConnell, C. R.; New Deal; Ownership vs. control; Samuelson, P. A.; Schumpeter, J. A.; Twentieth Century Fund

JEL Classifications

B31

A graduate at an early age of Harvard College and the Harvard Law School, Berle served in Army Intelligence in World War I and on the American delegation to the Paris Peace Conference, from which he emerged to denounce the terms of the Treaty, as did Keynes, though to a lesser audience. After practising law in New York, he joined the law faculty of Columbia University, where he became a member of the famous Brains Trust of Franklin D. Roosevelt. He was a close adviser of Roosevelt's, both before and after the latter's election to the Presidency.

In the later New Deal years, Berle served as an Assistant Secretary of State, then a senior position in the Department, and thereafter as ambassador to Brazil. In the years following World War II, he was chairman of the Liberal Party in New York and the long-time head of the Twentieth Century Fund, a foundation engaged in the active sponsorship of research in economic and social issues.

Berle's major contribution to economics, made in 1932 in conjunction with Gardiner C. Means in *The Modern Corporation and Private Property*, was in showing that authority in the modern large business enterprise moves ineluctably away from the owners of property to the managers and that by the time of research for the book the process was already far advanced. As a conclusion for conventional economics this, it is not too much to say, ranked in inconvenience with that of Keynes. Ownership no longer conveyed power in the great enterprise. Profit maximization was now by managers, not on behalf of themselves but for others largely unknown or, in pay and perquisites, for the managers themselves. Berle's conclusions also denied the independent, self-motivated, heroic role of the entrepreneur as offered in conventional economics, notably by Schumpeter.

Berle's contribution came from outside the conventional boundaries of the profession – from, of all things, a lawyer. Perhaps for this reason its importance was discounted, even denied, by many economists. In recent times, however, the truth of Berle's contentions has been recognized as personal profit maximization of managers – salaries, diverse perquisites, stock options, golden parachutes – has become one of the accepted scandals of the time. Nonetheless, Berle's role as one of the major innovating figures in economics has never been adequately recognized. In his textbook Paul Samuelson acknowledges *The Modern Corporation* as a classic; in Campbell R. McConnell's *Economics*, the most

widely used text in the United States, Berle's name does not even appear.

In his later years Berle returned in a perceptive and informative way to the subject of power, though not with the innovative force of his earlier work.

Selected Works

1932. (With G.C. Means.) *The modern corporation and private property.* New York: The Commerce Clearing House.
1959. *Power without property: A new development in american political economy.* New York: Harcourt, Brace & World.
1963. *The american economic republic.* New York: Harcourt, Brace & World.
1969. *Power.* New York: Harcourt, Brace & World.

Bernácer, Germán (1883–1965)

Mauro Boianovsky

Abstract

Bernácer contributed to macroeconomics the concept of 'disposable funds' and a new theory of interest. A lag between received and disbursed income underlies his view that aggregate equilibrium in the goods market emerges only if the amount of disposable funds is the same at the beginning and at the end of the period. Bernácer also argued that the rate of interest was determined outside the production system by land purchases and sales in the assets market. Economic fluctuations are decided by oscillations in the amount of disposable funds determined by the interaction between the markets for goods and for old assets.

Keywords

Aggregate equilibrium; Bernácer, G.; Böhm–Bawerk, E. von; Business cycles; Cash-in-advance constraint models; Crowding out; Disposable funds; Economic geography; Economic integration; Effective demand; George, H.; Interest rate determination; IS–LM model; Keynes, J. M.; Liquidity constraints; Money; Natural rate and market rate of interest; Physiocracy; Robertson, D.; Saving–investment equality; School of Salamanca; Speculative markets; Stabilization policies; Stocks and flows; Turgot, A. R. J

JEL Classifications

B31

Bernácer was born in Alicante, Spain, on 29 June 1883, and died on 22 May 1965 in the same city. He may be regarded as the first major monetary economist in the Spanish language since the School of Salamanca in the 16th century. Bernácer completed his studies at the Alicante School of Commerce (Escuela Superior de Comercio de Alicante) in 1901, where he was awarded the chair of industrial physics (Tecnología Industrial) in 1905. In that same year he started working on his big book *Sociedad y Felicidad – Ensayo de Mecánica Social,* which shows the influence of his physics background in the study of the economic aspects of social life, especially his distinction between the 'static and dynamics of wealth' in the study of 'social problems' such as business cycles and unemployment. That book was eventually published in 1916, some time after a study tour of eight months that had taken him to several European countries in 1911. In the next ten years, some of the main ideas presented in incipient form in *Sociedad y Felicidad* were further developed in two publications by Bernácer. His 1922 essay introduced into the economic literature the concept of 'disposable funds' ('disponibilidades') and its implications for the treatment of the demand for money and monetary dynamics. Bernácer sent 150 copies of that essay (with a French summary) to prominent economists and journals around the world. His 1925 book advanced a new approach to the origins and determination of interest as a variable decided outside the production system.

In the early 1930s Bernácer moved to Madrid to become the first director of the Research Service of the Bank of Spain. His appointment was probably influenced by his long 1929 article about the determination of the exchange rate as an equilibrium variable, in which he discussed in detail how to stabilize the external and internal values of the Spanish peseta and the conditions for returning to the gold standard system. He continued to teach, this time as professor of physics and chemistry at the School of High Commercial Studies of Madrid (Escuela de Altos Estudios Mercantiles). In 1940 long extracts from Bernácer's 1922 article were translated into English and published in *Economica* with a commentary by Dennis Robertson, who had been one of the recipients of that article in the 1920s. Robertson's article made Bernácer known to the Anglo-Saxon world and led him to restate the main theoretical and methodological features of his approach to monetary economics in a volume published in 1945. In the 1950s he wrote his last two books, dealing with economic integration and economic geography (1953) and summing up his views about economic dynamics and economic reform (1955). At about this time Bernácer retired from both his appointments as professor in Madrid and as director of research at the Bank of Spain.

Period Analysis and Disposable Funds

Bernácer's main contribution to economics is his analysis of the role played by money in the determination of economic variables such as income, employment, the rate of interest and the rate of exchange. He introduced the concept of a lag between received and disbursed income, which provided the starting-point of his discussion of aggregate disequilibrium in the market for goods. Bernácer's lag probably influenced the well-known Robertsonian related lag between received and disposable income. It follows from his concept of disposable funds (A) held at the beginning of the economic period, which, when added to the income (R) received during the period, give the upper limit of effective demand ($A + R$). Money balances are functionally classified into three grades, from minimum to maximum degree of disposability: (a) money demand by families to meet consumption; (b) money demand by businessmen for the conduct of their enterprises; and (c) new savings which have not yet been put by their owners to remunerative employment. Bernácer used the phrase 'disponibilidades' to refer to the last two classes. In order to determine the flow of 'effective demand' (D) it is necessary to subtract from A the amount of disposable funds left at the end of the period (A'), which gives the equation $R + (A - A') = D$, or, since R is identical with output P, the equation $P + (A - A') = D$. The last equation indicates that there is aggregate equilibrium (in the sense that production is equal to effective demand and the output produced is sold at the expected price) if the amount of disposable funds is the same at the beginning and at the end of the period ($\Delta A = 0$). The key to Bernácer's monetary economics is his notion that the spending decisions of economic agents (firms and families alike) in any given period of time are constrained by the amount of money they possess at the outset of that period. Bernácer was probably the first to introduce the main elements of what would become known in the literature as the 'cash-in-advance constraint' models developed in the 1960s.

Bernácer's approach to the business cycle was based on his distinction between the market for goods ('circulación productiva'), which decides the price level, and the market for 'valores de renta' or income-yielding assets ('circulación especulativa' or 'circulación financeira'), where the rate of interest is determined. Similar distinctions between aggregate markets for flows and stocks respectively would be deployed later in macroeconomic models put forward by John Hicks (IS–LM model), James Tobin and others. The interplay between those two markets explains fluctuations in income and employment in Bernácer's framework. The use of disposable funds to buy 'valores de renta' in the financial or speculative market does not change the condition of disposable funds, as they remain disposable in the hands of the sellers of assets. On the other hand, the use of disposable funds to purchase consumption goods and new capital goods brings about a change in their degree of disposability, as

they are turned into money income of the individuals involved in the production of goods. This constitutes 'effective demand', as opposed to 'potential demand' that does not involve a change in liquidity. Aggregate equilibrium can now be also described by the equality between saving and investment, which is the case if the saving flow is not directed to the purchase of 'valores de renta'. Economic fluctuations result from the opposite effects on the price level and the rate of interest of changes in disposable funds. When $\Delta\Delta A$ is negative in the upswing, prices of consumption goods are higher than anticipated and, since wages and salaries are temporarily fixed, employers will see their 'residual profits' increase. The ensuing stimulus to production and employment will cease when, under the impact of an increasing shortage of disposable funds in the 'speculative market', the rate of interest rises and saving is gradually directed to that market. This way, ΔA becomes positive, which explains the upper turning point of the business cycle. During the downswing, unanticipated falling prices bring about losses, which contributes (together with the constraint represented by a reduction of firms' liquidity) to a contraction in production and employment. The depression is characterized by widespread 'forced [or involuntary] unemployment' ('paro forsozo'), which is not solved by money-wage reductions, since lower wages will bring about a further fall in consumption demand and ensuing price reductions.

The Speculative Market and the Rate of Interest

The main factor in Bernácer's account of the business cycle is not the variability of investment demand by entrepreneurs, but the savers' decisions on how to allocate their disposable funds – purchase of new capital goods in the goods market or of old assets in the speculative market. The banking and credit system is incidental to Bernácer's framework, which is different from the well-known Wicksellian distinction between the 'natural' and the 'market' rates of interest. Bernácer's explanation of macroeconomic disequilibrium is based on another sort of divergence, that is, on differences between the rate of interest decided by the expected rate of return on new capital goods on one side, and the rate of interest determined by the relative yields of 'valores de renta' in the speculative market. The notion that the rate of interest is determined outside the system of current production is a crucial feature of the Bernácerian theoretical system. He argued that the rate of interest is determined not by the scarcity of capital goods as such, but by the scarcity of disposable funds. Moreover, given the identity between aggregate income and output, the rate of interest cannot be determined simply by saving and investment: if the disposable funds were used only to purchase the current output (of consumption and capital goods), the saving flow would necessarily be identical with the output of new capital goods, with no scarcity of funds in that market. The rate of interest can be positive only if a scarcity of disposable funds comes about because of the possibility of employing them outside the production system, that is, in the speculative market. The problem of the origin and determination of interest, according to Bernácer, consists in the search for an asset able to yield a 'free' rent without any production costs. He found it in land (in the broad sense of agricultural and urban land, as well as mines), not because of its productivity, but because it has a price and is exchangeable for other assets through money. In particular, the rate of interest is the determined variable in the equation relating its value to the price and the rent of land. Land, however, is not capital, and its purchase is not a real investment, since money remains disposable; hence, Bernácer explained how land's ability to produce rent is transmitted to other applications of money – especially to new capital goods – through the equilibrium between the marginal rates of return of old and new assets in the market. Such a mechanism, however, cannot work if the rate of return of investment in new capital goods falls to zero or below (which, of course, cannot happen to land and other income-yielding assets) in the depression, as pointed out by Bernácer. After he had put forward the main elements of his interest theory in 1916, Bernácer noticed several similarities with what Böhm-Bawerk used to call Turgot's 'fructification

theory' of interest, but observed that, in contrast with Turgot's, his approach was not based on the Physiocratic framework.

Bernácer would claim, after the publication of Robertson's article in 1940, that the dynamic approach to monetary economics introduced in his 1922 essay was the source of Robertson's own formulation of period analysis in 1926 and, via Robertson, of the 'fundamental equations' of Keynes's 1930 *Treatise on Money*. Whereas there are some grounds to substantiate Bernácer's claim, it should be noted that the economic policy conclusions he drew from his theoretical framework are far apart from those advocated by Robertson or Keynes. Bernácer was critical of attempted stabilization policies of both fiscal and monetary sorts, because of the crowding out effect and of the (destabilizing) impact of monetary and credit changes on prices. Instead, he believed that the market economy was an essentially efficient institution, except for the existence of the speculative market for income-yielding assets that kept the economy in a chronic state of unemployment. Bernácer's suggested solution was to make the amount of disposable funds constant by suppressing that market through the legal prohibition of the sale of land, which would bring the rate of interest to zero. Although this is somewhat reminiscent of Henry George's reform proposals in the 19th century, it should be noted that Bernácer supported neither George's tax reform nor George's approaches to economic fluctuations and the determination of interest. It is likely that Bernácer's idiosyncratic ideas about economic reform, as well as his rejection of macroeconomic stabilization policies, contributed to distracting interest from the depth of his economic theory and to explaining its relative lack of influence in Spain throughout his lifetime.

See Also

- ▶ George, Henry (1839–1897)
- ▶ Robertson, Dennis (1890–1963)
- ▶ Spain, Economics in
- ▶ Turgot, Anne Robert Jacques, Baron de L'Aulne (1727–1781)

Selected Works

1916. Sociedad y Felicidad – Ensayo de Mecánica Social. Madrid: F. Beltrán.
1922. La teoría de las disponibilidades como interpretación de las crisis económicas y del problema social. *Revista Nacional de Economía* 40: 535–562.
1925. *Interés del Capital*. Alicante: Lucentum.
1926. El ciclo económico. *Revista Nacional de Economía* 66 and 67: 3–30 and 155–179.
1929. La técnica del retorno al patrón oro (I, II, III). *Revista Nacional de Economía* 83, 84 and 85: 3–15, 223–239 and 405–418.
1945. *La Doctrina Funcional del Dinero*. Madrid: Consejo Superior de Investigaciones Cientificas.
1950. Money and freedom. *Kyklos* 4(2/3): 123–139.
1953. La Doctrina del Gran Espacio Económico. Madrid: Aguilar.
1955. Una Economía Libre, sin Crisis, y sin Paro. Madrid: Aguilar.
2005. En Torno a la Obra y Figura del Economista Germán Bernácer-Selección de Escritos y Conferencias. Alicante: Caja de Ahorros del Mediterráneo.

Bibliography

Almenar, S. 1999. Keynes's economic ideas in Spain before the *General Theory:* Spread, 'anticipations' and parallels. In *The impact of keynes on the economics in the 20th century*, ed. L. Pasinetti and B. Schefold. Cheltenham: Elgar.
Boianovsky, M., H. Dar, J. Presley, and P. Brañas-Garza. 2006. Cambridge and the Spanish connection: The contribution of Germán Bernácer. *History of Political Economy* 38: 407–436.
Freyre, J. 1957. Review of Bernácer (1955). *American Economic Review* 47: 436–438.
Pozuelo y Barnuevo, J. 1965. Germán Bernácer (1883–1965). *Revue d'Économie Politique* 75: 1231–1235.
Robertson, D. 1926. *Banking policy and the price level*. London: P.S. King & Son Ltd.
Robertson, D. 1940. A Spanish contribution to the theory of fluctuations. *Economica* 7(February): 50–65.
Ruiz, G. 1984. *Germán Bernácer – Un Economista Anticipativo*. Madrid: Pirámide.
Savall, H. 1975. *G. Bernácer – L'Héterodoxie en Science Économique*. Paris: Dalloz.
Wallich, H. 1947. Review of Bernácer (1945). *Review of Economic Statistics* 29: 65–66.

Bernoulli, Daniel (1700–1782)

S.L. Zabell

Keywords

Bernoulli, D.; Bernoulli, N.; Cramer, G.; D'Alembert, C.; Euler, L.; Goldbach, C.; Laplace, P. S.; Logarithmic utility; Maximum likelihood; Moral expectation; Probability; St Petersburg paradox; Utility

JEL Classifications

B31

Swiss mathematician and theoretical physicist; born at Groningen, 8 February 1700; died at Basel, 17 March 1782.

Daniel Bernoulli was a member of a truly remarkable family which produced no fewer than eight mathematicians of ability within three generations, three of whom–James 1 (1654–1705), John 1 (1667–1748) and Daniel – were luminaries of the first magnitude.

Although initially trained in medicine, in 1725 Daniel Bernoulli accepted a position in mathematics at the newly founded Imperial Academy in St Petersburg, but returned to Basel in 1733, holding successively the chairs in anatomy and botany, physiology (1743), and physics (1750–77). He was elected to membership in all of the major European learned societies of his day, including those of London, Paris, Berlin and St Petersburg, and maintained an extensive scientific correspondence which included both Euler and Goldbach.

Original in thought and prolific in output, Bernoulli worked in many areas but his most important contributions were to the fields of mechanics, hydrodynamics and mathematics. He enjoys with Euler, his close friend from childhood, the distinction of having won or shared no fewer than ten times the annual prize of the Paris Academy. His masterpiece, the *Hydrodynamica* (1738), contains a derivation of the *Bernoulli equation* for the steady flow of a non–viscous, incompressible fluid, and the earliest mathematical treatment of the kinetic theory of gases, including a derivation of Boyle's Law.

Bernoulli also made important contributions to probability and statistics, including an early application of the method of maximum likelihood to the theory of errors and an investigation of the efficacy of smallpox inoculation (Todhunter 1865, ch. 11). Nevertheless, his best-known contribution to this subject is unquestionably his 1738 paper 'Specimen theoriae novae de mensura sortis', which discusses utility, 'moral expectation' and the St Petersburg paradox.

The St Petersburg paradox (so called because Bernoulli's paper appeared in the *Commentarii* of the St Petersburg Academy) concerns a game, first suggested by Nicholas Bernoulli (Daniel's cousin) in correspondence with Montmort: a coin is tossed n times until the first head appears; 2^n ducats are then paid out. Paradoxically, the mathematical expectation of gain is infinite although common sense suggests that the fair price to play the game should be finite.

Bernoulli proposed that the paradox could be resolved by replacing the mathematical expectation by a moral expectation, in which probabilities are multiplied by personal utilities rather than monetary prices. Arguing that incremental utility is inversely proportional to current fortune (and directly proportional to the increment in fortune), Bernoulli concluded that utility is a linear function of the logarithm of monetary price, and showed that in this case the moral expectation of the game is finite.

Strictly speaking, Bernoulli's advocacy of logarithmic utility did not 'solve' the paradox: if utility is unbounded, then it is always possible to find an appropriate divergent series. Nor was he the first to adopt such a line of attack; the Swiss mathematician Gabriel Cramer had earlier written to Nicholas Bernoulli in 1728, noting that if utility were either bounded or proportional to the square root of monetary price, then the moral expectation would be finite. But it was via Bernoulli's paper that the utility solution entered the literature, and despite initial (and eccentric) criticism by D'Alembert, by the 19th century most treatises on probability would contain a section on moral expectation and the paradox.

An English translation of Bernoulli's 1738 paper on the St Petersburg paradox was published in *Econometrica* 22 (1954), 23–36, and is reprinted in *Precursors in Mathematical Economics*: *An Anthology,* ed. W.J. Baumol and S.M. Goldfeld, Series of Reprints of Scarce Works on Political Economy, No. 19, London: London School of Economics and Political Science, 1968, pp. 15–26. An English translation of Bernoulli's paper on maximum likelihood estimation appears in *Biometrika* 48 (1961), 1–18.

For further biographical information about Daniel Bernoulli and a detailed scientific assessment of his work, see the article by Hans Straub in *Dictionary of Scientific Biography,* vol. 2 (1970). The DSB also contains excellent entries on several other members of the Bernoulli family. Eric Temple Bell's *Men of Mathematics* (1937) contains a spirited, if not necessarily reliable, account of the Bernoullis.

Todhunter (1865, ch. 11) is still valuable as a summary of Bernoulli's work in probability; Todhunter's book is, as Keynes justly remarked, 'a work of true learning, beyond criticism'. For further information on Bernoulli's contributions to probability and statistics, see also Sheynin (1970, 1972) and Maistrov (1974, pp. 106–7, 110–18). The dispute with D'Alembert is discussed by Baker (1975, pp. 172–5); see also Pearson (1978, pp. 543–55, 560–65) and Daston (1979, pp. 259–79).

Useful discussions of Bernoulli's paper on the St Petersburg paradox include Leonard J. Savage (1954, pp. 91–5) and J.M. Keynes (1921, pp. 316–20). The mathematician Abel once wrote that one should read the masters and not the pupils; those who wish to follow Abel's advice will find challenging but rewarding Laplace's discussion of moral expectation in his *Théorie analytique des probabilités* (1812, ch. 10: 'De l'espérance morale').

The literature on the St Petersburg paradox up to 1934 is surveyed in Karl Menger (1934); an English translation of Menger's paper appears in M. Shubik (ed., 1967). For a discussion of the St Petersburg paradox in the context of an axiomatization of utility and probability other than that of Ramsey and Savage, see Jeffrey (1983, pp. 150–5). The paradox still continues to inspire interest and analysis; a recent example is Martin-Lof (1985).

Bibliography

Baker, K.M. 1975. *Condorcet: From natural philosophy to social mathematics*. Chicago: University of Chicago Press.

Bell, E.T. 1937. *Men of mathematics*. New York: Simon & Schuster; Harmondsworth: Pelican Books, 1953.

Bernoulli, D. 1738. *Hydrodynamica, sive de viribus et motibus fluidorum commentarii* …. Strasbourg: Johann Reinhold Dulsseker.

Daston, L.J. 1979. D'Alembert's critique of probability theory. *Historia Mathematica* 6: 259–279.

Jeffrey, R.C. 1983. *The logic of decision*. 2nd ed. Chicago: University of Chicago Press.

Keynes, J.M. 1921. *A treatise on probability.* London: Macmillan.

Laplace, P.S. 1812. *Théorie analytique des probabilités.* 2nd ed. Paris, 1814; 3rd ed, 1820.

Maistrov, L.E. 1974. *Probability theory: A historical sketch*. New York: Academic Press.

Martin-Lof, A. 1985. A limit theorem which clarifies the 'Petersburg paradox'. *Journal of Applied Probability* 22: 634–643.

Menger, K. 1934. Die Unsicherheitsmoment in der Wehrtlehre. *Zeitschrift für Nationalökonomie* 5: 459–485. Trans. in *Essays in mathematical economics in honor of Oskar Morgenstern*, ed. M. Shubik. Princeton: Princeton University Press, 1967.

Pearson, K. 1978. *The history of statistics in the 17th and 18th centuries*. New York: Macmillan.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Sheynin, O.B. 1970. Daniel Bernoulli on the normal law. *Biometrika* 57: 99–102.

Sheynin, O.B. 1972. D. Bernoulli's work on probability. *RETE Strukturgeschichte der Naturwissenschaften* 1: 273–300.

Straub, H. 1970. Bernoulli, Daniel. In *Dictionary of scientific biography*, ed. C.C. Gillispie, vol. 2, 136–146. New York: Scribner's.

Todhunter, I. 1865. *A history of mathematical theory of probability from the time of Pascal to that of Laplace*. Cambridge: Cambridge University Press, repr. New York: Chelsea, 1961.

Bernoulli, James [Jakob, Jacques] (1654–1705)

A. W. F. Edwards

Bernoulli was born in Basel on 27 December 1654 and died there on 16 August 1705, a scion of a famous family of Swiss mathematicians. In 1687

he was appointed Professor of Mathematics in the University of Basel, and besides major contributions to probability theory he made advances in the calculus, the theory of series, and mechanics.

In the field of probability his *Ars conjectandi* was published posthumously in 1713. Part I is a commentary, with text, on Huygens's *De ratiociniis in aleae ludo* of 1657, in the course of which Bernoulli gave the expression for the binomial distribution for general chances. For this reason 'binomial trials' are sometimes called 'Bernoulli trials', although in fact De Moivre published the expression earlier. Part II is *The doctrine of permutations and combinations*, written in ignorance of Pascal's *Traité du triangle arithmétique* and therefore not as novel as Bernoulli thought. Part III applies the theory of Part II to games of chance, while Part IV contains the celebrated limit theorem in probability in which Bernoulli derived an expression for the number of binomial trials required to ensure that the proportion of successes falls within stated limits with a certain specified probability. As the number of trials is increased, this probability tends to 1. He applied this theorem to the estimation of the binomial parameter, revealing a clear understanding of the problem of statistical estimation and thus inaugurating a continuing debate about the proper solution.

Bernstein, Eduard (1850–1932)

Tom Bottomore

Keywords

Bernstein, E.; Class; Class conflict; Engels, F.; Kautsky, K.; Marx, K. H.; Social democracy; Socialism; Weber, M.; Winstanley, G.

JEL Classifications

B31

Born in Berlin, 6 January 1850; died in Berlin, 18 December 1932. The son of a Jewish railway engineer and the seventh child in a large family of 15 children, Bernstein grew up in a lower middle-class district of Berlin in 'genteel poverty'. He did not complete his studies at the Gymnasium, and in 1866 he began an apprenticeship in a Berlin bank. Three years later he became a bank clerk and remained in this post until 1878, but he continued to study independently and for a time aspired to work in the theatre. He became a socialist in 1871, largely through sympathy with the opposition of Bebel, Liebknecht and others to the Franco–Prussian war, and strongly influenced by reading Marx's study of the Paris Commune, *The Civil War in France* (1871). In 1872 Bernstein joined the Social Democratic Workers' Party, and in 1875 he was a delegate to the conference in Gotha which brought about the union of that party with Lassalle's General Union of German Workers to form a new Socialist Workers' party, later the Social Democratic Party (SDP). From that time Bernstein became a leading figure in the socialist movement, and in 1878, just before Bismarck's anti-Socialist law was passed, he moved to Switzerland as secretary to a wealthy young socialist, Karl Höchberg, who expounded a form of utopian socialism in the journal *Die Zukunft* which he had founded. It was in 1878 also that Bernstein read Engels's *Anti-Dühring,* which, he said, 'converted me to Marxism', and he corresponded with Engels for the first time in June 1879. After some misunderstandings with Marx and Engels, who were suspicious of his relationship with Höchberg, Bernstein won their confidence during a visit to London and in January 1881, with their support, he became editor of *Der Sozialdemokrat* (the newspaper of the SDP, established in 1879). It was, as Gay (1952) notes, 'the beginning of a great career'.

In 1888 the Swiss government, under pressure from Germany, expelled Bernstein and three of his colleagues on the *Sozialdemokrat*, and they moved to London to continue publication there. The period of exile in England, which lasted until 1901, was crucial in the formation of Bernstein's ideas. He became a close friend of Engels, who made him his literary executor (jointly with Bebel), and developed a stronger interest in historical and

theoretical subjects, contributing regularly to Kautsky's *Die Neue Zeit* and publishing in 1895 his first major work, a study of socialism and democracy in the English revolution (entitled *Cromwell and Communism* in the English translation). Bernstein's major contributions in this study, which he later described as 'the only large scale attempt on my part to discuss historical events on the basis of Marx's and Engels's materialist conception of history', were to analyse the civil war as a class conflict between the rising bourgeoisie and both the feudal aristocracy and the workers, and to give prominence to the ideas of the radical movements in the revolution (the Levellers and Diggers), and in particular those of Gerrard Winstanley, who had been ignored by previous historians.

At the same time Bernstein established close relations with the socialists of the Fabian Society and came to be strongly influenced by their 'gradualist' doctrines and their rejection of Marxism. In a letter to Bebel (20 October 1898) he described how, after giving a lecture to the Fabian Society on 'What Marx really taught', he became extremely dissatisfied with his 'well-meaning rescue attempt' and decided that it was necessary 'to become clear just where Marx is right and where he is wrong'. Soon after Engels's death Bernstein began to publish in *Die Neue Zeit* (from 1896 to 1898) a series of articles on 'problems of socialism' which represented a systematic attempt to revise Marxist theory in the light of the recent development of capitalism and of the socialist movement. The articles set off a major controversy in the SDP, in which Kautsky defended Marxist orthodoxy and urged Bernstein to expound his views in a more comprehensive way, as he then proceeded to do in his book on 'the premisses of socialism and the tasks of social democracy' (1899; entitled Evolutionary Socialism in the English translation), which made him internationally famous as the leader of the 'revisionist movement'.

Bernstein's arguments in *Evolutionary Socialism* were directed primarily against an 'economic collapse' theory of the demise of capitalism and the advent of socialism, and against the idea of an increasing polarization of society between bourgeoisie and proletariat, accompanied by intensifying class conflict. On the first point he was attacking the Marxist orthodoxy of the SDP, expounded in particular by Kautsky, rather than Marx's own theory, in which the analysis of economic crises and their political consequences was not fully worked out, and indeed allowed for diverse interpretations (Bottomore 1985). The central part of Bernstein's study, however, concerned the changes in class structure since Marx's time, and their implications. In this view, the polarization of classes anticipated by Marx was not occurring, because the concentration of capital in large enterprises was accompanied by a development of new small and medium-sized businesses, property ownership was becoming more widespread, the general level of living was rising, the middle class was increasing rather than diminishing in numbers, and the structure of capitalist society was not being simplified, but was becoming more complex and differentiated. Bernstein summarized his ideas in a note found among his papers after his death: 'Peasants do not sink; middle class does not disappear; crises do not grow ever larger; misery and serfdom do not increase. There *is* increase in insecurity, dependence, social distance, social character of production, functional superfluity of property owners' (cited by Gay 1952, p. 244).

On some points Bernstein was clearly mistaken. With the further development of capitalism, peasant production has declined rapidly and has been superseded to a great extent by 'agri-business'; economic crises did become larger, at least up to the depression of 1929–33. It was his analysis of the changing class structure which had the greatest influence, becoming a major issue in the social sciences, and above all in sociology, in part through the work of Max Weber, whose critical discussion of Marxism in his lecture on socialism (1918) largely restates Bernstein's arguments. There is a more general sense in which Bernstein's ideas have retained their significance; namely, in their assertion of the increasingly 'social character' of production and the likelihood of a gradual transition to socialism by the permeation of capitalist society with socialist institutions. In a different form the same notion is expressed by Schumpeter (1942) in his conception of a gradual 'socialization of the economy'; a conception which can also be traced back to Marx (Bottomore 1985).

One other aspect of Bernstein's thought should be noted. Influenced by the neo-Kantian movement in German philosophy and by positivism (in an essay of 1924 he noted that 'my way of thinking would make me a member of the school of Positivist philosophy and sociology') Bernstein made a sharp distinction between science and ethics and went on to argue, in his lecture 'How is scientific socialism possible?' (1901), that the socialist movement necessarily embodies an ethical or 'ideal' element: 'It is something that *ought* to be, or a movement towards something that *ought* to be.' From this standpoint he criticized in a more general way a purely economic interpretation of history, and especially the kind of 'economic determinism' that was prevalent in the orthodox Marxism of the SDP; but in so doing he cannot be said to have diverged radically from the conceptions of Marx and Engels (and indeed he cited Engels's various qualifications of 'historical materialism' in support of his own views).

Bernstein's book met with a vigorous and effective response in Rosa Luxemburg's *Sozialreform oder Revolution* (1899), and the SDP became divided between 'radicals', 'revisionists', and the 'centre' (represented by Bebel and Kautsky); and although the latter retained control Bernstein remained a leading figure in the party until 1914. But his growing opposition to the war led him to form a separate organization in 1916 and then to join the left-wing Independent Social Democratic Party of Germany (USPD) in 1917. After the war Bernstein became increasingly disillusioned with the ineffectualness of the SDP in countering the reactionary nationalist attacks on the Weimar Republic, his influence waned, and his last years were spent in isolation.

See Also

► Social Democracy

Selected Works

1895. *Cromwell and communism.* London: Allen & Unwin, 1930.
1899. *Evolutionary socialism.* New York: Huebsch, 1909. Reprinted, New York: Schocken, 1961.
1901. Wie ist wissenschaftlicher Sozialismus möglich? *Sozialistische Monatshefte.*

Bibliography

Bottomore, T. 1985. *Theories of modern capitalism*. London: Allen & Unwin.
Gay, P. 1952. *The dilemma of democratic socialism*. New York: Columbia University Press.
Luxemburg, R. 1899. *Sozialreform oder revolution.* Trans. as *Reform or revolution*. New York: Three Arrows, 1937.
Schumpeter, J.A. 1942. *Capitalism, socialism and democracy.* 5th ed. London: Allen & Unwin. 1976.
Weber, M. 1918. Socialism. English Trans. In *Max weber the interpretation of social reality,* ed. J.E.T. Eldridge. London: Michael Joseph, 1970.

Bibliographic Addendum

A recent biography is Steger, M. *The quest for evolutionary socialism: Eduard Bernstein and social democracy.* Cambridge: Cambridge University Press, 1997.

Berry, Arthur (1862–1929)

J. K. Whitaker

A Cambridge mathematician who dabbled briefly in economics, Berry was born on 28 May 1862 in Croydon and died on 15 August 1929 in Cambridge. Entering King's College, Cambridge, in 1881, he was Senior Wrangler in the Mathematical Tripos of 1885 and became a Fellow of King's in 1886. After extension lecturing, he returned permanently to Cambridge in 1889. Thereafter, apart from administering Cambridge extension lecturing from 1891 to 1895, he devoted himself to King's and the teaching of mathematics, highly regarded but publishing little.

Berry's social and political interests were broad. As an undergraduate he had co-founded the Cambridge Economic Club, to which he delivered a paper on factory legislation (Berry 1886). He must have attended Alfred Marshall's

lectures and subsequently, at the latter's request, lectured on mathematical economics from 1891 to 1900, after which W.E. Johnson took over. Marshall also instigated Berry's only two publications on economic theory (Berry 1891a, b). The first, which survives only as an abstract, was a significant contribution to the emerging marginal productivity theory of distribution on lines already sketched by Marshall. The second was a masterful resolution of a dispute between Marshall and F.Y. Edgeworth over the theory of barter, background letters on which are reproduced by Guillebaud (1961, Vol. II, pp. 791–8). After 1891 Berry drifted away from economics, partly because of heavy administrative work, and partly because of friction with Marshall over the question of women's status at Cambridge.

As an economist (and also more generally) Berry was talented but without a strong drive towards original work. His best-known publication was a history of astronomy for extension audiences (Berry 1898). See *The Times* (1929) for further biographical information.

Selected Works

1886. *Factory legislation.* Text of a paper presented to the Cambridge Economic Club, Cambridge, privately printed.

1891a. The pure theory of distribution. *Report of the Sixtieth (1890) Meeting of the British Association for the Advancement of Science.* Reprinted in *Precursors in mathematical economics,* ed. W.J. Baumol and S.M. Goldfeld. London: London School of Economics, 1968.

1891b. Alcune brevi parole sulla teoria del baratto. *Giornale degli Economisti,* June.

1898. *A short history of astronomy.* London: Murray.

References

Guillebaud, C.W. (ed.). 1961. *Alfred Marshall: Principles of economics,* 9th (Variorum) ed. London: Macmillan.

The Times. London. 1929. Obituary: Arthur Berry, 19 August.

Bertalanffy, Ludwig von (1901–1972)

Kenneth E. Boulding

Primarily a biologist, Bertalanffy is recognized as the father of General Systems Theory and a founder of the Society for General Systems Research. Born near Vienna in 1901 he taught at the University of Vienna (1934–48), the University of Ottawa (1948–54), the University of Alberta (1961–9) and the State University of New York at Buffalo (1969–72). Like many pioneers, his work was recognized during his own lifetime by only a few, but his influence continues to grow. His work, especially on the theory of open systems, led the way to a more unified theory of organisms and organizations stretching from the biological to all the social sciences. He was an important contributor to what might be called the 'post-Newtonian' movement in the sciences, rejecting the reductionism of logical positivism, insisting that systems have hierarchies of complexity, each with its own patterns and methods, allowing for indeterminacy, recognizing that equilibrium is unknown in the real world except as an approximation, and stressing the generality of both ontogenetic and phylogenetic processes.

Main-line economics has remained solidly Newtonian, and the influence of Bertalanffy and of General Systems has been very small. Nevertheless the growing interest in evolutionary models and in more organic approaches to the growth and structure of firms suggest that the hope expressed by Alfred Marshall that economics could learn much from biology, and Veblen that economics might become an evolutionary science, may indicate a future somewhat different from the past. In such a case the importance of Bertalanffy's contribution will be more fully recognized.

See Also

► General Systems Theory

Bertrand Competition

Michael R. Baye and Dan Kovenock

Abstract

This article presents the classic Bertrand model of oligopolistic price competition and shows how alternative assumptions on economic primitives – such as the structure of demand and cost functions, tie-breaking rules, and product differentiation – shape Nash equilibrium prices and profits. We also discuss the related Bertrand–Edgeworth model of price competition in which consumers may be rationed – either strategically or due to capacity constraints – and illustrate how alternative rationing rules influence equilibrium.

Keywords

Bertrand competition; Bertrand equilibrium; Bertrand paradox; Bertrand, J. L. F.; Bertrand–Edgeworth competition; best response (reply); capacity; Cournot, A. A.; duopoly; Edgeworth cycles; homogeneous products; mixed-strategy equilibria; monopoly; Nash equilibrium; oligopoly; price competition; product differentiation; pure-strategy equilibria; rationing rules; residual monopoly price; strategic complements; supermodularity; winner-take-all

JEL Classifications

D4

'Bertrand competition' refers to a model of oligopoly in which two or more firms compete by simultaneously setting prices and in which each firm is committed to provide consumers with the quantity of the firm's product they demand given these 'posted prices'. The concept is named after the French mathematician Joseph Louis François Bertrand (1822–1900) who, in an 1883 (Bertrand 1883) review of Cournot (1838), was critical of Cournot's use of quantity as the strategic variable in his famous duopoly model of market rivalry. In his critique, Bertrand described how, in Cournot's duopoly environment where identical firms produce a homogeneous product under a constant unit cost technology, price competition would lead to price undercutting and a downward spiral of prices. Bertrand erroneously reasoned that this process would continue indefinitely, thereby precluding the existence of an equilibrium. It is now widely recognized that an equilibrium exists not only in Bertrand's original formulation but in a plethora of other environments in which firms sell either homogeneous or differentiated products.

Formally, *Bertrand competition* is a normal form game in which each of $n \geq 2$ players (firms), $i = 1,2,\ldots,n$, simultaneously sets a price $p_i \in P_i = [0, \infty)$. Under the assumption of profit maximization, the payoff to each firm i is

$\pi_i(p_i, p_{-i}) = p_i D_i(p_i, p_{-i}) - C_i(D_i(p_i,p_{-i}))$, where p_{-i} denotes the vector of prices charged by all firms other than i, $D_i(p_i, p_{-i})$ represents the total demand for firm i's product at prices $(p_i\, p_{-i})$, and $C_i(D_i(p_i, p_{-i}))$ is firm i's total cost of producing the output $D_i(p_i, p_{-i})$. A *Bertrand equilibrium* is a Nash equilibrium of this game; that is, a vector of prices $\left(p_i^*, p_{-i}^*\right)$ such that, for each player i, $\pi_i\left(p_i^*, p_{-1}^*\right) \geq \pi_i\left(p_i, p_{-1}^*\right)$ for all $p_i \in P_i$.

The Bertrand Paradox

In the 'classic' model of Bertrand competition, each of the n firms produces an identical product at a constant unit cost of c; that is, $C_i(q_i) = cq_i$. Since their products are perfect substitutes, firms effectively compete for the total demand, $D(p)$, that a monopolist serving the entire market would obtain by pricing at p. The firm setting the lowest price gets all of this demand; in the event of a tie, the firms charging the lowest price share total demand equally. Total demand is sufficiently well-behaved to ensure that the corresponding monopoly profit function, $\pi(p) \equiv pD(p) - C(D(p))$, is not only continuous, but (a) has a unique maximizer, the monopoly price p^M; (b) satisfies

$\pi(p) < \pi(c) = 0$ for $p < c$; and satisfies (*c*) $0 < \pi(p) < \pi(p^M) < \infty$ for all $p^M > p > c$. Despite the continuity of $\pi(p)$, each firm faces a discontinuous profit function

$$\pi_i(p_i, p_{-i}) = \begin{cases} (p_i - c)D(p_i) & \text{if } p_i < p_j \text{ for all } j \neq i \\ (p_i - c)D(p_i)/m & \text{if } i \text{ ties } m-1 \text{ other firms for low price} \\ 0 & \text{otherwise} \end{cases}$$

because a firm that prices even slightly above the lowest price gets no demand. In this classic setting with 'well-behaved' demand and constant marginal cost, (p_i^*, p_{-i}^*) is a *Bertrand equilibrium* if and only if $p_j^* \geq c$ for every firm *j* and at least two firms set price equal to *c*. Consequently, all firms earn zero profits in equilibrium, a result that has come to be known as the *Bertrand paradox*. The paradox stems from the fact that, while a monopolist would earn strictly positive profits by charging a price in excess of marginal cost, it takes only two firms to completely dissipate the monopoly profits and achieve the competitive outcome. In a Bertrand equilibrium, all transactions take place at marginal cost (*c*), and all firms earn zero profits.

The proof of this proposition follows in part from the original intuition of Bertrand. Since the products are perfect substitutes, consumers will purchase only from a firm that charges the lowest price in the market, $p_L \equiv \min_j p_j$. First, $p_L \leq p^M$ in any equilibrium; otherwise, any firm could profitably deviate by lowering its price to p_M. Second, $p_L \geq c$ in any equilibrium; otherwise, a firm charging p_L (and thus earning strictly negative profits) could profitably deviate by increasing its price to *c*. Third, if $p^M \geq p_L > c$, then at least one firm could increase its profit by unilaterally undercutting p_L by a small amount. Hence, $p_L = c$ in any equilibrium. Fourth, if only a single firm charged a price of $p_L = c$, it would earn a payoff of zero, and could increase its price to $p' > c$ (but below the second-lowest price) to earn a positive profit. Thus, in any equilibrium at least two firms charge a price of $p_L = c$. Finally, since the only firms attracting any consumers are those pricing at $p_L = c$, all firms earn zero profits. Furthermore, no firm can unilaterally change its price to earn positive profits.

One consequence of this argument is that when $n = 2$ there is a unique Bertrand equilibrium in the classic model: both firms set the common price $p_1^* = p_2^* = c$. When $n > 2$, there is a unique symmetric equilibrium (in which $p_i^* = c$ for all *i*) and a continuum of asymmetric equilibria (where two or more firms price at *c* and one or more firms charge prices arbitrarily higher than *c*).

Although the Bertrand paradox result summarized above for the case of identical constant unit costs is stated in terms of pure strategies and a symmetric tie-breaking rule, the paradox also obtains for the extension of strategy spaces to allow for mixed-strategies as well as other tie-breaking rules. Alternative tie-breaking rules include 'winner-take-all sharing' (where a fair randomizing device is used to determine the identity of the firm that services the entire market in the event of a tie for the lowest price) and 'unequal sharing' (where firms tying for the lowest price receive an unequal fraction of total market demand in the event of a tie for the lowest price).

Baye and Morgan (1999) have shown that if the monopoly profit function, $\pi(p)$, is unbounded, there exists (in addition to the Bertrand paradox equilibria) a continuum of non-degenerate mixed strategy equilibria in which each firm earns positive profits. For instance, suppose market demand is given by $D(p) = p^{\alpha}$, where $\alpha \in (-\infty, -1/n)$ is the elasticity of market demand. In this case, one can show that there is a unique symmetric Cournot (quantity-setting) equilibrium in which each firm earns positive profits and the equilibrium market price is $p^* = [n\alpha/(1 + n\alpha)]c$. In contrast, under Bertrand competition any symmetric profit level $\pi^* \in (0, \infty)$ (including profit levels above the Cournot profit) can be achieved in an (atomless) symmetric mixed strategy equilibrium. Equilibrium mixed strategies that support these positive profit levels are described by the cumulative distribution

function $F(p) = 1 - \pi^*/\pi(p)$ on $[\pi^{-1}(\pi^*),\infty)$, where $\pi(p) = (p-c)p^{\alpha}$.

Even with a bounded monopoly profit function $\pi(p)$, the coexistence of positive profit equilibria and (zero profit) Bertrand paradox equilibria can arise for alternative cost functions and sharing rules. For instance, with a symmetric tie-breaking rule (see Dastidar 1995), if firms have identical cost functions that are increasing and strictly convex in output, a symmetric zero profit equilibrium may exist in which each firm prices at p^0, where p^0 satisfies $p^0D(p^0)/n–C(D(p^0)/n) = 0$. In addition, however, a continuum of positive profit symmetric pure-strategy equilibria can arise in which each firm charges a price contained in an interval above p^0. Intuitively, with strictly convex costs, a firm that deviates by undercutting such a price would increase its demand (and revenues) by a factor of n, but the firm's cost would increase by a factor greater than n.

This result for bounded demand and identical convex costs is based on a symmetric tie-breaking rule; with convex costs, different results generally obtain for other tie-breaking rules. For instance, under the winner-take-all tie-breaking rule (see Baye and Morgan 2002), any firm charging the price p_L earns a payoff of $\pi(p_L)/\#L$, where $\#L$ is the number of firms charging the price p_L. In this case, if $\pi(p_L) > 0$, some firm could gain by undercutting p_L by a small amount (a firm pricing above p_L could increase its payoff from zero to $\pi(p_L - \varepsilon) > 0$; a firm that tied another firm at p_L could increase its profits from $\pi(p_L)/\#L$ to $\pi(p_L - \varepsilon)$ by slightly undercutting p_L). Consequently, an argument similar to that for the case of constant unit costs implies that, with bounded demand and convex costs, any equilibrium under the winner-take-all sharing rule involves at least two firms charging a price p_L such that $\pi(p_L) = 0$, so that the (zero profit) Bertrand paradox is the only configuration of firm profits.

With bounded demand and identical concave costs, a similar argument reveals that any equilibrium under the winner-take-all sharing rule involves at least two firms charging a price p_L such that $\pi(p_L) = 0$ (Baye and Morgan 2002). However, under a symmetric sharing rule, concave costs (increasing returns) are problematic for the existence of a Bertrand equilibrium in either pure or mixed strategies. To illustrate, consider a duopoly in which market demand is given by $D(p) = 1 - p$ for $p \in [0, 1]$, and in which each firm has an identical concave cost function

$$C_i(q_i) = \begin{cases} 0 & \text{if} \quad q_i = 0 \\ f + cq_i & \text{if} \quad q_i > 0 \end{cases}$$

where $1 > c > 0$ and $f < [(1 - c)/2]^2$. Note that c represents marginal cost and f is a fixed cost that may be avoided by producing zero output. One may readily verify that a monopolist would earn strictly positive profits by pricing at the monopoly price $p^M = (1 + c)/2$, and that the minimum 'breakeven price' is $p^0 + [(1 + c) - [(1 - c)^2 - 4f]^{1/2}]2$; that is, $0 = \pi(p^0) > \pi(p)$ for all $p < p^0$. Under a winner-take-all sharing rule, $p_1 = p_2 = p^0$ is a pure-strategy Nash equilibrium and firms earn zero profits in this 'Bertrand paradox' equilibrium. In contrast, under a symmetric tie-breaking rule there does not exist an equilibrium (in pure or mixed strategies).

The intuition for the failure of existence of equilibrium with a symmetric tie-breaking rule in this example is as follows. Clearly, neither firm has an incentive to price below p^0 (since monopoly profits are negative for such prices and a firm can guarantee a payoff of zero by pricing at $p_i = 1$). If both firms priced at p^0 with probability one, symmetric sharing implies that they would earn negative profits, since $C_i(D(p^0)/2) > C_i(D(p^0))/2$. Thus, p^0 is strictly less than the upper bound of the support of at least one firm's (possibly degenerate) mixed strategy. Let $p^H > p^0$ denote highest of the upper bounds of the supports of the two firms' mixed strategies. In any equilibrium, at most one firm has a mass point at p^H; otherwise, there would be a positive probability of a tie at this price and a firm could gain by reallocating mass to lower prices. If there is a mass point at p^H, the firm charging p^H with positive probability must earn its equilibrium profits at this price, which are necessarily zero since it is undercut with certainty. If there is no mass point at p^H, then a firm whose support includes p^H must achieve its equilibrium payoff when pricing at p^H,

and since p^H is undercut with certainty, this equilibrium payoff is zero.

Therefore, at least one firm i whose support includes p^H earns an equilibrium payoff of zero. Moreover, since firm i earns an equilibrium payoff of zero, p^0 must be the upper bound of the support of the other firm j's mixed strategy; if the upper bound of j's support was $p' \in (p^0, p^H]$, firm i could increase its profits by reallocating probability mass to some price below p'. Thus, if there is an equilibrium, at least one firm must charge a price of p^0 with probability one. However, since firm i charges prices in the interval $[p^0, p^H]$ and not all mass is at p^0, it follows that there exists some price $p'' \in (p^0, p^H]$ such that firm j could gain by reallocating mass from p^0 to p'', a contradiction. Hence, there does not exist an equilibrium in pure or mixed strategies.

Bertrand–Edgeworth Competition

In an early critique of Bertrand and Cournot, Edgeworth (1925) observed that the Bertrand paradox may not obtain if firms are capacity constrained. Indeed, in the analysis above, if firm i's demand $D_i(p_i; p_{-i})$ is greater than firm i's largest competitive supply at p_i, $s_i(p_i) = \max\{\arg\max_q p_i q - c_i(q)\}$, then firm i would earn higher profits by supplying a quantity strictly less than that demand and rationing customers. A variant of Bertrand competition, known as 'Bertrand–Edgeworth competition', allows any firm to ration the demand that it faces at given prices by only providing its optimal or competitive supply at its price. Rationing may stem from a physical capacity constraint, k_i, that prevents firm i from producing more than k_i units (as in Edgeworth's original formulation), or more generally, from a firm's strategic incentive to refuse to fulfil the quantity demanded of all consumers at a given price. Under Bertrand–Edgeworth competition one must therefore specify how demand is rationed when a firm's quantity demanded at given prices exceeds the amount of product it produces.

Two prominent rationing rules used in this context are efficient rationing (in which case the good is first allocated to consumers who most highly value the product) and proportional rationing (in which case the good is allocated to a fraction of consumers without regard to their valuations of the product). In the duopoly case, for instance, efficient rationing means that if $p_i > p_j$; firm i's 'residual' demand is $D_i(p_1, p_2) \equiv \max\{0, D(p_i) - s_j(p_j)\}$. Under proportional rationing, firm i's demand is $D_i(p_1, p_2) \equiv \max\{0, D(p_i)[1 - s_j(p_j)/D(p_j)]\}$. Under both rationing rules, the firm charging the lowest price enjoys a demand of $D(p_j)$. It is typically assumed that, in the event of a tie, total demand is allocated in proportion to firms' competitive supplies; that is, if both firms charge a price of p, firm i gets a share $\alpha_i = s_i(p)/(s_1(p) + s_2(p))$.

For the special case of a duopoly in which each firm has a constant marginal cost (c) up to a capacity of k_i, the cost functions are:

$$C_i(q_i) = \begin{cases} cq_i & \text{if} \quad 0 \le q_i \le k_i \\ \infty & \text{if} \quad q_i > k_i \end{cases}$$

In this case, under the assumption of well-behaved demand, $s_i(p_i) = k_i$ for all $p_i \ge c$; that is, each firm opts for a 'corner solution' at full capacity when price exceeds marginal cost. Under both efficient and proportional rationing, if $D(c) \le k_i$, $i = 1, 2$, then neither firm's capacity constraint ever binds and the Bertrand paradox arises under the same conditions as set forth above; the unique equilibrium is $p_1^* = p_2^* = c$. Characterization of equilibrium when one or more firms is capacity constrained at a price equal to c depends on whether each firm is capacity constrained at its 'residual monopoly price' when its rival sets $p_j = D^{-1}(k_1 + k_2)$. The term 'residual monopoly price' refers to a firm's optimal price, given its capacity constraint and residual demand (the demand that remains after the other firm has sold its capacity). Note that, in equilibrium, neither firm would ever set a price below $D^{-1}(k_1 + k_2)$, for at such a price total demand exceeds total capacity, and a firm could increase its price without losing sales. Characterization of equilibrium when $D(c) > k_i$ for one or more firms then depends on whether $p_1 = p_2 = D^{-1}(k_1 + k_2)$ is an equilibrium.

If, for each firm i, $D^{-1}(k_1 + k_2)$ is the residual monopoly price when firm j sets $p_j = D^{-1}(k_1 + k_2)$, then $p_1^* = p_2^* = D^{-1}(k_1 + k_2)$ is the unique Bertrand–Edgeworth equilibrium. If some firm i's residual monopoly price exceeds $D^{-1}(k_1 + k_2)$ when $p_j = D^{-1}(k_1 + k_2)$, then the unique equilibrium is in non-degenerate mixed-strategies.

The residual monopoly price depends on the rationing rule. For proportional rationing, $D_i(p_1, p_2) \equiv \max\{0, D(p_i)[1 - k_j/D(p_j)]\}$ for any given p_j, and hence firm i's demand is proportional to $D(p_i)$. This implies that, ignoring firm i's capacity constraint, the residual monopoly price based on $D_i(p_1, p_2)$ corresponds to the standard monopoly price, $p^M = \arg\max_p\{(p - c)D(p)\}$. When $p_j = D^{-1}(k_1 + k_2) < p^M$, firm i has sufficient capacity to satisfy residual demand at p^M, and hence p^M is firm i's residual monopoly price; if $p_j = D^{-1}(k_1 + k_2) \geq p^M$, concavity of the monopoly profit function implies that $p_i = D^{-1}(k_1 + k_2)$ is firm i's residual monopoly price. It follows that, for proportional rationing, $p_1^* = p_2^* = D^{-1}(k_1 + k_2)$ is the unique Bertrand– Edgeworth equilibrium as long as $D^{-1}(k_1 + k_2) \geq p^M$.

Under efficient rationing, $D_i(p_1, p_2) \equiv \max\{0, D(p_i) - k_j\}$, so that ignoring firm i's capacity constraint, the residual monopoly price is $p_i^R = \text{argmax}_{pi}\{(p_i - c)\max(0, D(p_i) - k_j)\}$. It follows that $p_i^R < p^M$. When $p_j = D^{-1}(k_1 + k_2) < p_i^R$, firm i has sufficient capacity to satisfy residual demand at P_i^R, and hence p_i^R is firm i's residual monopoly price; if $p_j = D^{-1}(k_1 + k_2) \geq p_i^R$, concavity of the monopoly profit function implies that $p_i = D^{-1}(k_1 + k_2)$ is firm i's residual monopoly price. Hence, $D^{-1}(k_1 + k_2)$ is firm i's residual monopoly price when firm j sets $p_j = D^{-1}(k_1 + k_2)$ if and only if $D^{-1}(k_1 + k_2) \geq p_i^R$. This implies that the region in which a pure strategy equilibrium arises is larger for the case of efficient rationing than under proportional rationing. In fact, since the unconstrained residual profit-maximization problem faced by firm i under efficient rationing may be written in terms of either price or quantity, p_i^R is the price arising in a Cournot setting where firm i's output is a best response to an output of k_j by the rival. Hence, if k_j is less than or equal to firm i's Cournot best response to k_j, firm i is capacity constrained and its residual monopoly price equals $D^{-1}(k_1 + k_2)$. Consequently, $p_1^* = p_2^* = D^{-1}(k_1 + k_2)$ is the unique Bertrand–Edgeworth equilibrium when each firm's capacity is less than or equal to its Cournot best response (given unit cost c) to the other firm's capacity.

Outside of the above regions of capacity, the only Bertrand–Edgeworth equilibria are in non-degenerate mixed strategies in which firms randomize prices over a common interval of prices that exceed c and earn positive expected profits. This corresponds to the regions of capacities in which 'Edgeworth cycles' arise (Edgeworth 1925). As before, these mixed strategies depend on the rationing rule. For proportional rationing, these mixed strategies are generally difficult to derive; see Davidson and Deneckere (1986) for a characterization. For efficient rationing, these mixed strategies have been characterized by Kreps and Scheinkman (1983), and entail the firm with the larger capacity earning an expected payoff that equals the monopoly profit associated with the residual demand (with symmetric capacities, each firm earns this expected payoff). The firm with the larger capacity earns the higher payoff.

To summarize, only two types of pure-strategy equilibria exist under Bertrand–Edgeworth duopoly with constant unit cost. When capacity constraints do not bind, the classic Bertrand equilibrium arises and the unique equilibrium is for each firm to price at marginal cost to earn zero profits. When capacities are sufficiently small, firms price above marginal cost (at a price that clears all capacity) and earn positive profits in the unique Bertrand–Edgeworth equilibrium. When capacities are in an intermediate range, the equilibrium is generally unique, but in non-degenerate mixed strategies. Firms' prices exceed marginal cost with probability one, and firms earn positive profits.

Positive profit equilibria can also arise in homogeneous product Bertrand settings in which firms endogenously choose capacities. Specifically, consider a twostage game where, in the first stage, firms simultaneously commit to a capacity, and in the second stage firms simultaneously engage in Bertrand–Edgeworth competition. Under both

efficient and proportional rationing, capacity commitment in the first stage permits both firms to avoid the Bertrand paradox in the second stage to earn positive profits. Under efficient rationing, capacity choice followed by Bertrand–Edgeworth competition leads, under fairly general conditions, to equilibrium prices that are identical to those that would arise in a Cournot (quantity setting) duopoly where firms' unit costs are the sum of capacity and production costs; see Kreps and Scheinkman (1983) and Deneckere and Kovenock (1996). Under proportional rationing, the Cournot outcome arises only if per unit capacity costs are sufficiently large. Otherwise, equilibria may arise in which capacities are asymmetric and non-degenerate mixed strategies are played at the pricing stage; see Davidson and Deneckere (1986).

Product Differentiation

Bertrand competition with differentiated products is fundamentally different from Bertrand competition with homogenous products. With differentiated products, the demand for a firm's product is not generally discontinuous at p_L; a firm does not generally lose all of its demand by pricing slightly above p_L, nor does it steal all of rival firms' demands by pricing below p_L. In the classical model of differentiated-product Bertrand competition with downward sloping demands and costs that are non-decreasing in output, each firm's profit function, $\pi_i(p_i, p_{-i})$, is assumed to be twice continuously differentiable, with $\partial\pi_i/\partial p_i\partial p_j > 0$ (strategic complements) and $\partial^2\pi_i/\partial p_i^2 < 0$.

With suitable assumptions on firms' demands and costs, a Bertrand equilibrium, (p_i^*, p_{-i}^*), is simply the solution to the system of first-order conditions implied by each firm's profit-maximizing pricing decision:

$$\frac{\partial\pi_i(p_i^*, p_{-i}^*)}{\partial p_i} = 0 \text{ for all } i = 1, 2, \ldots, n.$$

Alternatively, one may use the implicit function theorem and use firm i's first-order condition to obtain firm i's optimal price as a function of the prices charged by the other firms: $p_i = \rho_i(p_{-i})$. The function ρ_i is called firm i's best-response (best-reply, reaction) function, and a Bertrand equilibrium in the case of differentiated products corresponds to the intersection of the firms' best-response functions. Total differentiation of firm i's first-order condition reveals that $d\rho_i/dp_j = -(\partial\pi_i/\partial p_i\partial p_j)/(\partial^2\pi_i/\partial p_i^2) > 0$; that is, strategic complementarities and the concavity of firm i's profits in p_i imply that firm i's best response function is upward sloping.

Notice that, at (p_i^*, p_{-i}^*),

$$\frac{\partial\pi_i(p_i^*, p_{-i}^*)}{\partial p_i} = [p_i^* - C_i'(D_i(p_i^*, p_{-i}^*))]\frac{\partial D_i(p_i^*, p_{-i}^*)}{\partial p_i} + D_i(p_i^*, p_{-i}^*) = 0.$$

Consequently, under mild regularity conditions firm i's equilibrium price exceeds its marginal cost. Furthermore, firms may charge different prices and earn positive profits in a differentiated product Bertrand equilibrium. These results may be extended to the case where $\pi_i(p_i, p_{-i})$ is not differentiable by appealing to the more general notion of supermodularity (Vives 1990; Milgrom and Roberts 1990) rather than strategic complementarity (Bulow et al. 1985).

For the duopoly case with linear demands and constant unit costs, strategic complementarity ($\partial\pi_i/\partial p_i\partial p_i > 0$) arises naturally when the duopolists' products are substitutes in consumption ($\partial D_i/\partial p_j > 0$). In this case the firms' best-response functions are not only upward sloping (as is implied by strategic complementarity) but linear; consequently, there is a unique Bertrand equilibrium (see Cheng 1985). Singh and Vives (1984) have shown that, in this linear duopoly case, even though each firm prices above its marginal cost in a differentiated-product Bertrand equilibrium, prices are lower under Bertrand competition than would arise in a differentiated-product Cournot (quantity setting) model. This result for linear demand and costs extends to markets with more than two firms

when all firms' products are substitutes in consumption (Häckner 2000).

See Also

- ▶ Cournot Competition
- ▶ Supermodularity and Supermodular Games

Bibliography

Baye, M.R., and J. Morgan. 1999. A folk theorem for one–shot Bertrand games. *Economics Letters* 65: 59–65.

Baye, M.R., and J. Morgan. 2002. Winner-take-all price competition. *Economic Theory* 19: 271–282.

Bertrand, J. 1883. (Review of) *Théorie Mathématique de la Richesse Sociale* par Léon Walras: *Recherches sur les Principes Mathématiques de la Théorie des Richesses* par Augustin Cournot. *Journal des Savants* 67: 499–508.

Bulow, J., J. Geanakoplos, and P. Klemperer. 1985. Multimarket oligopoly: Strategic substitutes and complements. *Journal of Political Economy* 93: 488–511.

Cheng, L. 1985. Comparing Bertrand and Cournot equilibria: A geometric approach. *RAND Journal of Economics* 16: 146–152.

Cournot, A. 1838. *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. Paris: Hachette.

Dastidar, K.G. 1995. On the existence of pure strategy Bertrand equilibrium. *Economic Theory* 5: 19–32.

Davidson, C., and R. Deneckere. 1986. Long-run competition in capacity, short-run competition in price, and the Cournot model. *RAND Journal of Economics* 17: 404–415.

Deneckere, R., and D. Kovenock. 1996. Bertrand–Edgeworth duopoly with unit cost asymmetry. *Economic Theory* 8: 1–25.

Edgeworth, F.Y. 1925. The pure theory of monopoly. *Papers Relating to Political Economy* 1: 111–142.

Häckner, J. 2000. A note on price and quantity competition in differentiated oligopolies. *Journal of Economic Theory* 93: 223–239.

Kreps, D.M., and J.A. Scheinkman. 1983. Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics* 14: 326–337.

Milgrom, P., and J. Roberts. 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica* 58: 1255–1277.

Singh, N., and X. Vives. 1984. Price and quantity competition in a differentiated duopoly. *RAND Journal of Economics* 15: 546–554.

Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19: 305–321.

Bertrand, Joseph Louis François (1822–1900)

Martin Shubik

Keywords

Bertrand, J. L. F.; Cournot, A. A.; Duopoly; Edgeworth, F. Y.; Mathematical economics; Tâtonnement; Walras, L.

JEL Classifications

B31

Bertrand was born and died in Paris. He was an eminent but not great mathematician, graduate and professor of mathematics at the Ecole Polytechnique and from 1862 to 1900 a member of the Collège de France. His relevance to economic thought comes in his criticism of 'pseudo-mathematicians' in the *Journal des Savants* (1883) where he reviewed *Théorie mathématique de la richesse sociale* of Walras and *Recherches sur les principes mathématiques de la théorie des richesses* of Cournot. It is doubtful if Bertrand considered the problems of formal economic modelling more than casually, viewing the two works through the eyes of a mathematician with little substantive interest or understanding. His comments on Cournot were not only somewhat harsh, but as the subsequent developments in oligopoly theory and the theory of games have shown, both Cournot's model of duopoly and Bertrand's remodelling of duopoly with price rather than quantity as a strategic variable are worth investigation. Cournot's model has been (until recently) more generally treated than Bertrand's model. It remained for Edgeworth to point out the limitations of Bertrand's model (see Shubik 1959). Bertrand also raised objections to the reference and realism of the process description of Walras of 'tâtonnement'.

It has been suggested (Blaug and Sturges 1983) that Bertrand's critical review was used by opponents of mathematical economics as the basis for

their position. Although explicit proof of this is hard to establish the tone and force of Bertrand's critique makes this highly probable.

Selected Works

1883. (Review of) Théorie mathématique de la richesse sociale par Léon Walras: Recherches sur les principes mathématiques de la théorie des richesses par Augustin Cournot. *Journal des Savants,* September: 499–508.

Bibliography

Blaug, M., and P. Sturges, eds. 1983. *Who's who in economics*. Brighton: Wheatsheaf Books.
Byron, G.H. 1899–1900. Joseph Bertrand. *Nature* 1591 (61): 614–616.
Shubik, M. 1959. *Strategy and market structure*. New York: Wiley.
Storick, D.L. 1970. *Joseph Louis François Bertrand. Dictionary of Scientific Biography*, vol. 2. New York: Scribners.

Bettelheim, Charles (Born 1913)

Peter Nolan

Bettelheim has been a life-long Marxist for whom the theory and practice of the transition to socialism has been the central object of analysis. He has written influential theoretical works (e.g. Economic Calculation and Forms of Property, The Transition to Socialist Economy, Studies in the Theory of Planning), as well as studies of the political economy of different countries. The most important of these are on India (1968) – he was a consultant to the Indian government during the development of its planning system in the 1950s; on China (1974) he has visited China several times; and on the USSR (1946, 1976, 1978) – he reads Russian and has researched on the Soviet Union since the 1930s. He was influenced deeply by the Chinese cultural revolution, which shed new light on his view of the 'transition to socialism'. He considered that China had broken decisively (and correctly) from the Soviet Union's 'state capitalist' path. In the USSR, argued Bettelheim (following Mao), primacy was given to the 'development of the productive forces' at the expense of attempting to transform the system of unequal 'production relations', which formed the 'objective basis for the existence of classes'. His account of the Maoist attempt to break down workplace inequalities of power, income and status struck a powerful chord among many Western socialists at a time when Stalinism was being increasingly questioned, when confidence was high in the possibility of moving rapidly towards socialism, and before the mainstream of Western socialism had swung towards Euro-Communism.

Selected Works

1946. *La planification sovietique*. Paris: Marcel Rivière.
1959. *Studies in the theory of planning*. Bombay.
1968. *India independent*. London: Macgibbon & Kee.
1974. *Cultural revolution and industrial organization in China*. New York: Monthly Review Press.
1975. *The transition to socialist economy*. Hassocks: Harvester Press.
1976. *Economic calculation and forms of property*. London: Routledge.
1976. *Class struggles in the USSR, 1917–1923*. New York: Monthly Review Press.
1978. *Class struggles in the USSR, 1923–1930*. New York: Monthly Review Press.

Beveridge Curve

Eran Yashiv

Abstract

The Beveridge curve depicts a negative relationship between unemployed workers and job vacancies, a robust finding across countries.

The position of the economy on the curve gives an idea as to the state of the labour market. The modern underlying theory is the search and matching model, with workers and firms engaging in costly search leading to random matching. The Beveridge curve depicts the steady state of the model, whereby inflows into unemployment are equal to the outflows from it, generated by matching.

Keywords

Beveridge curve; Beveridge, W. H.; Business cycle; Excess demand and supply; Frictions; Information costs; Job search; Matching function; Microfoundations; Phillips curve; Unemployment; Vacancies; Wage inflation

JEL Classifications

E24

The Beveridge curve depicts a negative relationship between unemployed workers (u) and job vacancies (v). The interest in the curve is related to the role it plays in aggregate models, which study labour market outcomes and dynamics. The position of the economy on the curve gives an idea as to the state of the labour market; for example, a high level of vacancies and a low level of unemployment would indicate a 'tight' labour market. The literature has attempted to explain the coexistence of unemployment and vacancies, their negative relationship, and the implied dynamics.

The curve is named after William Beveridge, a British lord, lawyer, head of academic institutions, Member of Parliament, and founder of the modern British welfare state. In a 1944 report (Beveridge 1944), Beveridge discussed the relationship between the demand for workers, captured by vacancies, and the rate of unemployment. While he did not plot a curve or present a table with a comparison of u and v, he offered detailed data on these variables and discussed them at some length. His analysis implied that there is a negative relationship between them. In this early work he tackled many of the issues that remain under study in this field: the potential mismatch between unemployed workers and job vacancies, aggregate demand factors versus reallocation factors (for example, deficient overall demand for labour as opposed to low demand in particular industries), trend versus cyclical changes (for example, changes in u and v along the business cycle versus long-run changes), and measurement issues (such as the various possible ways of mismeasuring vacancies).

The negative $u - v$ relationship is a robust finding across countries, though shifts of the curve over time are often observed. This can be seen, for example, in a 16-country graphical description of the curve presented in Layard et al. (2005, pp. 36–7). Detailed descriptions and analyses of the empirical findings concerning the Beveridge curve for the United States are to be found in Blanchard and Diamond (1989), and for the UK in Pissarides (1986).

What underlies this negative relationship? The early literature of the late 1950s and in the 1960s dealt with the curve in the context of exploring excess demand in the labour market and its influence on wage inflation. This was motivated by the extensive study of the Phillips curve that took place in those years. The literature typically defined excess demand as unfilled vacancies less unemployed workers, considered the data on these variables, and then looked at the relationship between measures of excess demand and wage behaviour. This literature recognized that, even when there is no excess supply, there is positive unemployment due to frictions. It derived a negatively sloped $u - v$ curve from a model of distinct labour markets, interacting at different levels of disequilibrium, with the markets at points off both labour supply and labour demand curves. The $u - v$ curve was shown to be stationary and observed u and v points were expected to cycle around it. Movements up and down the curve reflect increases and decreases in the excess demand for labour. The curve itself can shift as a result of changes in the speed of market clearing or changes in the sectoral composition of labour demand. The observed $u - v$ data may be a compound of structural shifts of the curve together with cyclical movements about it. Key contributions to this strand of work were progressively made by Dow and Dicks-Mireaux (1958), Lipsey (1960), Holt and David (1966), Hansen (1970), and Bowden (1980).

In the 1970s and 1980s an alternative approach was developed – the search and matching model. A key difference between this model and the early literature is its derivation of vacancies and unemployment as equilibria, rather than disequilibria, phenomena. The model was developed in the work of Peter Diamond, Dale Mortensen, and Christopher Pissarides (see Pissarides 2000, for a detailed exposition, and Yashiv 2006, for a recent survey). The model may be briefly described as follows. Workers and firms engage in costly search to find each other. Firms spend resources on advertising, on posting job vacancies, on screening and, subsequently, on training. Workers spend resources on job search, with costs pertaining to activities such as collecting information and applying for jobs. Workers and firms are assumed to be randomly matched. After matching, the worker and the firm engage in bilateral bargaining over the wage. The matching process assumes frictions such as informational or locational imperfections. It is formalized by a 'matching function' that takes searching workers and vacant jobs as arguments and produces a flow of matches (m), and is given by $m = m(u, v)$. It is continuous, nonnegative, increasing in both its arguments, and concave. Typically, it is assumed to be constant returns to scale. The flow into unemployment results from job-specific shocks to matches that arrive at the Poisson rate λ. These shocks may be explained as shifts in demand or productivity shocks. Once a shock arrives, the firm closes the job down. The evolution of the unemployment rate ($\dot{u}$) is therefore given by the difference between the separation flow (λ times the employment rate $1 - u$) and the matching flow:

$$\dot{u} = \lambda(1 - u) - m(u, v). \tag{1}$$

Denote the rate at which workers are matched to jobs (the job finding rate) by $p = \frac{m}{u}$ so that $m = pu$. In the steady state the rate of unemployment is constant, so setting $\dot{u} = 0$ the following obtains:

$$u = \frac{\lambda}{\lambda + p}. \tag{2}$$

This is the Beveridge curve: as p depends on m, it depends on both u and v, and this equation can be represented in vacancy (v) – unemployment (u) space by a downward-sloping curve. The mechanism is the following. When vacancies v rise, matching m rises, and so the job finding rate p rises. Workers find jobs at a faster rate and unemployment u declines. Vacancies themselves are determined by a firm optimality equation, equating vacancy costs and benefits at the margin.

As can be seen in the equations above, the matching function plays a crucial role in generating the Beveridge curve. Petrongolo and Pissarides (2001) provide a comprehensive survey of estimation of this function, finding the following main features: (*a*) the prevalent specification is Cobb–Douglas, that is, $m = \mu u^{\alpha} v^{\beta}$; (*b*) usually constant returns to scale ($\alpha + \beta = 1$) is found, though some studies have produced evidence in favour of increasing returns to scale; (*c*) many studies have added other variables – such as demographical or geographical variables, incidence of long-term unemployment, and UI – finding some of them significant, but not changing the preceding findings; (*d*) these general patterns are robust across countries and time periods.

Research along the lines of this model – in progress – is likely to provide a richer account of the Beveridge curve: the matching function is studied for microfoundations, heterogeneity is explicitly explored, endogenous separations are allowed for, interactions with capital investment are considered, and learning and on- the-job search leading to job-to-job movements are incorporated. Going beyond this strand of the literature, research is also beginning to explore equilibrium search models, which feature a Beveridge curve, with alternative $u - v$ meeting processes, not modelled as matching functions. Thus, the Beveridge curve remains a topic of active research in macroeconomics and labour economics, more than 60 years after it was first studied.

See Also

▶ Beveridge, William Henry (1879–1963)

Bibliography

Beveridge, W. 1944. *Full employment in a free society.* London: George Allen and Unwin.
Blanchard, O., and P. Diamond. 1989. The Beveridge curve. *Brookings Papers on Economic Activity* 1: 1–60.
Bowden, R. 1980. On the existence and secular stability of the u-v loci. *Economica* 47: 35–50.
Dow, J., and L. Dicks-Mireaux. 1958. The excess demand for labour: A study of conditions in Great Britain, 1946–56. *Oxford Economic Papers* 10: 1–33.
Hansen, B. 1970. Excess demand, unemployment, vacancies and wages. *Quarterly Journal of Economics* 84: 1–23.
Holt, C., and M. David. 1966. The concept of vacancies in a dynamic theory of the labor market. In *Measurement and interpretation of job vacancies*, ed. NBER. New York: Columbia University Press.
Layard, R., S. Nickell, and R. Jackman. 2005. *Unemployment: Macroeconomic performance and the labour market*. 2nd ed. Oxford: Oxford University Press.
Lipsey, R. 1960. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1862–1957: A further analysis. *Economica* 27: 1–31.
Petrongolo, B., and C. Pissarides. 2001. Looking into the black box: A survey of the matching function. *Journal of Economic Literature* 39: 390–431.
Pissarides, C. 1986. Unemployment and vacancies in Britain. *Economic Policy* 1: 499–559.
Pissarides, C. 2000. *Equilibrium unemployment theory.* 2nd ed. Cambridge, MA: MIT Press.
Yashiv, E. 2006. Labor search and matching in macroeconomics. Institute for the Study of Labour (IZA) Discussion Paper no. 2743.

Beveridge, William Henry (1879–1963)

Jose Harris

Keywords

Beveridge, W. H.; Keynesian analysis of unemployment; Mixed economy; Pigou, A. C.; Poor Law; Redistribution of income; Robbins, L. C.; Social insurance; Transfer payments; Unemployment; Welfare economics; Welfare state

JEL Classifications

B31

Beveridge is chiefly remembered as a social and administrative reformer, whose *Social Insurance and Allied Services* (1942) set out the basic principles and structure of the post-war welfare state. Paradoxically, however, he thought of himself chiefly as an academic economist whose significance for posterity would lie in the fields of manpower policy and the theory of prices. Throughout his life his approach to economic problems was resolutely inductive and empirical, in contrast with the deductive and analytical method characteristic of most English economists. His early work, *Unemployment*: *A Problem of Industry* (1909), was based on detailed statistical analysis of the case-papers of applicants for unemployment relief. It drew attention to the structural, geographical and informational barriers that stood in the way of a perfect market for labour; and although its challenge to orthodox theory was practical rather than theoretical, it helped to erode belief in a natural economic equilibrium. Later editions of *Unemployment* (revised with the help of Lionel Robbins) were more strongly influenced by classical economic thought, but Beveridge never abandoned his belief that unemployment could only be cured by state intervention to organize and rationalize the market for labour. Beveridge in the 1930s was initially highly critical of the Keynesian analysis of unemployment; and although during the early 1940s he gradually absorbed many aspects of Keynesian thought, his *Full Employment in a Free Society* (1944) differed markedly from Keynes in its emphasis on the need for physical as well as fiscal controls over the economy and, in particular, on manpower planning.

Beveridge's early work on unemployment convinced him that there was a close and measurable connection between levels of economic activity and movements of prices. In the early 1920s he embarked upon what he came to see as his life's work; namely, the compilation of historical and statistical data relating to movements of prices since the 12th century. Beveridge's data convinced him that unemployment was caused, both nationally and internationally, by falls in the prices of primary products (though he failed to consider the possibility that the sequence of causation might lie

in the other direction). Beveridge's resistance to the use of analytical models meant that his data was of limited value to (and indeed often mocked by) economic theorists. Since his death, however, his material has been a seam of gold to many economic historians. Only one volume of the proposed project was ever published, *Prices and Wages in England from the Twelfth to the Nineteenth Century,* vol. I (1939), but much unpublished material survives among Beveridge's papers in the British Library of Political Science and the Institute of Historical Research.

Although Beveridge is often seen as a leading protagonist of the 'mixed' economy, his writings on economic policy displayed a recurrent scepticism about how far it was possible to reconcile state intervention with consumer sovereignty. His study of *British Food Control* (1928) suggested that there were advantages and disadvantages in both a 'laissez faire' and a 'command' economy, but that it was both logically and practically impossible to have the two in combination. Such doubts were partially allayed by the transformation of popular attitudes which appears to have occurred during the Second World War, but were never fully resolved. In his writings on social welfare, Beveridge appears to have been little influenced by, and indeed largely unconscious of, the growing body of contemporary writings on welfare economics produced by theorists like Pigou. His approach to social insurance, and to transfer payments generally, was that of an early 19th-century utilitarian, modified by a sociological and humanitarian perspective. All his proposals on social security display a concern to maintain some of the central economic tenets of the Poor Law (maintenance of incentives, encouragement to private saving, strict avoidance of relief-in-aid-of-wages) together with more 'organic' goals such as national efficiency and the maintenance of civilized minimum standards. His arguments for or against various methods and degrees of 'redistribution' were nearly always rooted in pragmatism or rule-of-thumb propositions about human behaviour, rather than in rigorous marginal analysis. Even in the most collectivist and 'socialistic' period of his career, he was insistent that claims to welfare should be rooted as far as possible in 'contract' rather than 'status'. His general perception of social welfare should be seen as that of a popular political theorist rather than that of an academic economist; though clearly his ideas in this field were both influenced by, and had wider implications for, economic thought.

Selected Works

1909. *Unemployment*: *A problem of industry.* London: Longmans & Co. Another ed., *Unemployment*: *A problem of industry, 1909 and 1930.* London: Longmans & Co., 1930.
1928. *British food control.* London: Oxford University Press.
1931. *Tariffs*: *The case examined.* London: Longmans & Co. Popular ed., 1932.
1936. *Planning under socialism, and other addresses.* London: Longmans & Co.
1939. *Prices and wages in England, from the twelfth to the nineteenth century,* vol. 1. London: Longmans & Co.
1942. *Social insurance and allied services*: *The Beveridge report in brief.* London: HMSO.
1944. *Full employment in a free society*: *A report.* Long: G. Allen & Unwin. 2nd ed., 1960.
Also numerous articles in *Economic Journal, Economica, Sociological Review, Politica,* and elsewhere.

Bias Correction

Jinyong Hahn

Keywords

Asymptotic theory; Bias correction; Empirical likelihood; Generalized method of moments; Limited information maximum likelihood; Nuisance parameters; Panel models; Two-stage least squares

JEL Classifications
C11; C13

Bias correction is a statistical technique used to remove the bias of an estimator. An unbiased estimator is such that its expectation is equal to the parameter of interest. Many introductory statistics textbooks discuss the desirability of having an unbiased estimator, although it is quickly pointed out that unbiasedness alone cannot be a good criterion for an estimator. This is usually illustrated by comparing two estimators with the use of a concrete loss function, where it is noted that an unbiased estimator with a large variance may be inferior to a biased estimator with a small variance.

Analysis of exact finite sample theory is difficult, or impossible, for many estimators. Therefore, sampling properties of econometric estimators are usually discussed in the context of asymptotic approximation. Many estimators used in econometrics are consistent and asymptotically efficient, so the bias is usually a non-issue in such first-order asymptotic theory. On the other hand, the first-order asymptotic theory may fail to provide a good approximation to the exact finite sample distribution of an estimator, and even an asymptotically unbiased estimator may have a significant bias under small sample sizes. Higher-order asymptotic approximation may then be used to understand the finite sample properties, including the approximate bias. To be more specific, suppose that we use an estimator $\hat{\theta}$ to estimate the parameter of interest θ_0. For many cases, $\hat{\theta}$ allows a three term stochastic expansion

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=\hat{T}_1+\hat{T}_2/\sqrt{n}+\hat{T}_3/n+O_p\left(n^{-3/2}\right),$$

where n is the sample size. The higher-order asymptotic bias of $\hat{\theta}$ is given by $b0/n$, where

$$b_0=\lim_{n\to\infty}E\left[\hat{T}_2\right].$$

In the recent literature, bias correction is usually understood to be a method of removing such approximate bias b_0/n. These methods include analytical corrections such as the standard textbook expansion for functions of sample means, and the more complicated formulas required for other estimators. They also include jackknife and bootstrap bias corrections. Correction of approximate bias is usually accompanied by increase of variance, and early literature such as Pfanzagl and Wefelmeyer (1978) focused on the efficiency aspects of bias correction. In general, bias correction cannot be always advocated on efficiency grounds.

Bias correction has received renewed attention in the more recent literature. When there are many nuisance parameters, the parameters of interest are typically estimated with significant biases. The biases are often so severe that removal of such biases almost always results in efficiency gain. Two strands of literature deal with models with many nuisance parameters. First, when a parameter of interest is estimated with many instruments, the resultant estimator may be quite biased. For example, the two-stage least squares estimator (2SLS) tends to be severely biased when there are many first-stage coefficients to be estimated; see for example Bekker (1994). It has been noted that some estimators are not sensitive to the presence of such nuisance parameters, and the instrumental variables literature is focused on developing such robust estimators. For linear simultaneous equations models, the limited information maximum likelihood estimator (LIML) was shown to have very little bias for linear models. For nonlinear models, it was shown that the empirical likelihood (EL) estimator tends to be less biased than the generalized method of moments estimator (GMM) when there are many moment restrictions; see Newey and Smith (2004).

The second strand of literature in which bias correction has played an important role is concerned with panel models. Parameters of interest in panel models are usually estimated with substantial bias when fixed effects are estimated; see Neyman and Scott (1948). The literature examined methods of removing such bias. Hahn and Newey (2004) proposed that the bias be estimated and subtracted from the estimator itself.

Arellano (2003) and Woutersen (2002) proposed that the moment equation be modified.

See Also

► Two-Stage Least Squares and The *K*-Class Estimator

Bibliography

Arellano, M. 2003. Discrete choices with panel data. *Investigaciones Económicas* 27: 423–458.

Bekker, P. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62: 657–681.

Hahn, J., and W. Newey. 2004. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72: 1295–1319.

Newey, W., and R. Smith. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72: 219–255.

Neyman, J., and E. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16: 1–31.

Pfanzagl, J., and W. Wefelmeyer. 1978. A third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis* 8: 1–29.

Woutersen, T. 2002. Robustness against incidental parameters. Working Paper No. 20028. Department of Economics/University of Western Ontario.

Biased and Unbiased Technological Change

Peter L. Rousseau

Abstract

This article provides working definitions of biased and unbiased technological change based on the relative responses of the marginal products of capital and labour that occur in the face of economic shocks. These Hicksian definitions are distinguished from others that focus on how technology augments the production function. The bias and augmentation of technical progress are then linked through the substitutability of labour and capital. Examples of 'labour-biased' and 'capital-biased' technological change from the 19th century to the present illustrate these ideas.

Keywords

Biased and unbiased technological change; CES production function; Cobb–Douglas functions; Elasticity of substitution; Information technology; Neutral production functions; Skill-bias; Technical change

JEL Classifications

O3

Among the central problems in growth economics is how to organize thinking about technological progress and its role in macroeconomic outcomes. In *The Theory of Wages* (1932), John Hicks offered a set of classifications for technical change that remains in common use. These classifications are based on the observation that inventions are unlikely to increase the marginal products of all factors of production in the same proportion, but rather will affect the marginal products of some factors more than others. Take, for example, the baseline two-factor neoclassical production function:

$$Y = F(K, L), \tag{1}$$

where Y is aggregate output, K is the capital stock, and L is labour. One way to introduce a technology parameter A is to place it at the front of the production function as

$$Y = AF(K, L). \tag{2}$$

Notice that A enters linearly, so that a doubling of the technology parameter also doubles output. Technological progress of this type is said to be 'unbiased' or 'Hicks neutral' in that the ratio of the marginal products of capital and labour used in the production process does not change. In this case, progress simply requires a renumbering of production isoquants.

Innovations are rarely neutral, however, and for this reason economists have naturally been more interested in cases where technological change alters the ratio of marginal products. When this occurs, technological change is said to be 'biased'. Hicks defines the bias as 'labour-saving' when the marginal product of capital increases more than that of labour for a given capital–labour ratio, thereby increasing the demand for capital. 'Capital-saving' technical progress occurs when the marginal product of labour rises more than that of capital for a given capital-labour ratio, thereby increasing the demand for labour. Nowadays economists simply refer to technological change that is labour-saving in the Hicksian sense as having a 'capital bias,' and change that is capital-saving in the Hicksian sense as having a 'labour bias.' This avoids confounding the bias of a given technological change with the way that it enters the production function.

An alternative concept proposed by R.F. Harrod (1937, 1948) defines technological change as neutral if the marginal product of capital is unchanged at a given capital–output ratio. Another way of stating this is that, under a constant rate of interest and an infinite supply of capital at that rate, a technological change is 'Harrod-neutral' if it leaves the length of the production process unaltered. H. Uzawa (1961) shows that this implies a production function of the form

$$Y = F(K, AL), \tag{3}$$

where AL is a unit of 'effective' labour. Note that this formulation is *not* neutral in the Hicksian sense unless the production function is Cobb–Douglas. Economists commonly refer to (3) as a 'labour-augmenting' production function, but it does not follow that technological change is necessarily labour-biased in the Hicksian sense of relative marginal products.

The opposite symmetric case to Harrod-neutrality defines an invention as neutral if the wage rate remains unchanged at a constant labour-output ratio. This implies a production function of the form

$$Y = F(AK, L), \tag{4}$$

where AK is a unit of 'effective' capital. Economists often refer to this 'capital-augmenting' form of the production function as 'Solow-neutral,' but only because Robert Solow (1959) was first to use this form to model technological progress. Once again, this formulation is *not* neutral in the Hicksian sense unless the production function is Cobb–Douglas, and changes in A are not necessarily capital-biased in the Hicksian sense. R. Sato and M.J. Beckmann (1968) offer a useful taxonomy of these and other 'neutral' production functions.

Of the three output equations shown above, it turns out that only the second (that is, labour-augmenting) form is consistent with a settling down to constant growth under steady technological progress and assumptions of constant returns to scale and diminishing marginal rates of substitution in production. Thus, if we are interested in neoclassical models that move beyond Cobb–Douglas production and possess a steady state, it is useful for technology to multiply labour and make it more effective. Since US wages have risen over the past century while the rental rate has remained relatively steady, the labour-augmenting formulation is at least a priori consistent with the evidence from the United States.

To distinguish technological progress that is factor-augmenting from their underlying Hicksian factor-biases, it is necessary to consider the elasticity of substitution between the factors as technical change occurs. Daron Acemoglu (2002) illustrates this with a CES (that is, constant elasticity of substitution) production function of the form.

$$Y = \left[w(A_L L)^{\frac{\sigma-1}{\sigma}} + (1-w)(A_K K)^{\frac{\sigma-1}{\sigma}} \right] \frac{\sigma}{\sigma - 1}, \tag{5}$$

where σ is the elasticity of substitution between capital and labour, A_L and A_K are factor-specific technology parameters, and w is a weight ($0 \leq w \leq 1$) that measures the relative importance of each factor. The factors are gross substitutes when $\sigma > 1$, whereas they are gross complements

when $\sigma < 1$. With $\sigma > 1$, substitutability between factors allows both the augmentation and bias of technological change to lean towards the same factor. In the case where $\sigma < 1$, however, a capital-augmenting technological change (or a rise in A_K) actually increases demand for the complementary input (that is, labour) more than it increases the demand for capital. The excess demand for labour raises its marginal product more than that of capital, leading to a labour bias in production. Similarly, a labour-augmenting technological change (or a rise in A_L) leads to a capital-bias when $\sigma < 1$. When $\sigma = 1$ the production function is Cobb–Douglas and an increase in A does not produce a bias towards either factor.

Hicks and A.C. Pigou (1920) have contended that most technological change is capital biased, and the American experience in the latter half of the 19th century would seem to support this view. Innovations such as the Bessemer process of steel-making, new distillation methods in petroleum refining, and the adoption of European reduction methods in flour milling, as noted by John James (1983), led to capital deepening and economies of scale in these industries that increased concentration. Such technological changes seem so important that the rise of big business around the turn of the 20th century is sometimes attributed to them. Though this view probably overstresses the role of technology in the evolution of industrial structure over this period, it is interesting that the capital bias observed in industries for which the story fits were a result of labour augmentation (that is, a rise in A_L) and inelastic factor substitution (that is, $\sigma < 1$).

Electrification offers another example. Prior to its arrival, manufacturing had been designed around the rigidities of steel shafts that ran through the length of a factory and were turned in unison by a single water or steam-powered generator. Afterwards, as Warren Devine (1983) describes, the organization of work gradually evolved to exploit the open factory structure that electric unit drive made possible. Unit drive meant less time spent maintaining complex systems of leather straps and pulleys that transferred power from the rotating steel shafts to the machines, and less down time caused by the need to stop all production to repair a single machine. Electrification and unit drive also made it economical for factories to stay open longer. These innovations made labour more productive (that is, raising A_L), but more focused machinery also reduced the amount of labour that was needed to operate a factory ($\sigma < 1$), raising the marginal product of capital more quickly than that of labour and producing a capital bias. The bias leaned even more towards capital as the diffusion of electricity began to mature, and labour-saving innovations such as vacuum cleaners, toasters, and electric blast furnaces became commonplace.

But is the apparent capital-bias in technological change largely 'induced' by changes in factor prices? Charles Kennedy (1964) points out that falling capital prices will motivate individuals to build more inventions that economize on labour than they would build at constant factor prices. Since the prices of capital goods have declined fairly consistently for more than a century and a half, it seems natural that the vast majority of induced inventions would have been capital biased. At the same time, it is important to distinguish biased technological progress (that is, an outward movement and shift along an isoquant) from movements along a fixed isoquant that arise from changes in factor prices, since such changes do not represent technological progress at all. Noting these potential biases, Hicks concludes that 'autonomous' inventions, meaning those not prompted by decline of a relative factor price, need not be predominantly capital biased. Indeed, information technology (IT) presents an example where the bias may have moved in the opposite direction.

Computers reduced expenditures on specialized and/or mechanical office machines, thereby making capital more productive (that is, raising A_K). At the same time, labour also became more productive as skilled individuals learned how to use computers to perform complex tasks and less-skilled individuals accomplished routine tasks much more quickly (that is, raising A_L). Thus, there seem to be complementarities between IT and skilled workers, raising the return to skill and producing a 'skill bias', while there has been some substitution of computers for less skilled

individuals, pressing towards a capital bias. On the whole, however, the complementarity effects so far have outweighed substitution effects, leading to a labour bias. As an invention in the method of inventing, IT has also led to a wide range of induced innovations, both capital- and labour-saving. Design tools used by engineers, for example, have improved the quality of capital goods and allowed more new products to be created. The availability of a broad base of knowledge on the World Wide Web from all over the globe has also transmitted the information needed to make labour more productive.

Is IT typical of the type of technological change that is likely to continue, starting with a labour bias but spawning new innovations that are for the most part labour-saving? If so, parsing out the components of labour bias, and particularly understanding the role of skill bias in the post-war US economy, seems at the core of understanding the role that technology will play in 21st century economic growth.

See Also

- ▶ Hicks, John Richard (1904–1989)
- ▶ Skill-Biased Technical Change
- ▶ Technical Change

Bibliography

Acemoglu, D. 2002. Directed technical change. *Review of Economic Studies* 69: 781–809.

Devine, W.D. Jr. 1983. From shafts to wires: Historical perspective on electrification. *Journal of Economic History* 43: 347–372.

Harrod, R.F. 1937. Review of Joan Robinson's essays in the theory of employment. *Economic Journal* 47: 326–330.

Harrod, R.F. 1948. *Towards a dynamic economics*. London: Macmillan.

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

James, J.A. 1983. Structural change in American manufacturing, 1850–1890. *Journal of Economic History* 43: 433–459.

Kennedy, C. 1964. Induced bias in innovation and the theory of distribution. *Economic Journal* 74: 541–547.

Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.

Sato, R., and M.J. Beckmann. 1968. Neutral inventions and production functions. *Review of Economic Studies* 35: 57–66.

Solow, R.M. 1959. Investment and technical change. In *Mathematical methods in the social sciences*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Palo Alto: Stanford University Press.

Uzawa, H. 1961. Neutral inventions and the stability of growth equilibrium. *Review of Economic Studies* 28: 117–124.

Bickerdike, Charles Frederick (1876–1961)

John S. Chipman

Keywords

Acceleration principle; Aftalion, A.; Bickerdike, C. F.; Depreciation; Durability of capital; Edgeworth, F. Y.; Elasticities approach to the balance of payments; Frisch, R. A.; Gestation period of capital; Hansen, A.; Incipient tariff; Johnson, H. G.; Kahn, R. F.; Lerner, A. P.; Local public finance; Metzler, L.A.; Optimal tariffs; Price discrimination; Robinson, J. V

JEL Classifications

B31

Bickerdike was born in England (whereabouts unknown) on 15 May 1876 and died in Wallington, Surrey, on 3 February 1961. He studied at Oxford from 1895 to 1899 where he received his BA degree in 1899 and MA in 1910. Upon winning the Cobden Prize for an essay summarized in Bickerdike (1902) he became a protégé of Edgeworth. After serving briefly as Lecturer on Economics and Commerce at the University of Manchester (1910–1912) he entered the civil service with a position in the Board of Trade, where he remained until his retirement in 1941.

Bickerdike's published work consists of 15 articles and 38 book reviews, all (save two of the

articles) in the *Economic Journal*. He is chiefly known as the originator of the theory of incipient and optimal tariffs (1906, 1907), according to which a country can always gain by imposing a sufficiently small tariff on its imports and can maximize its welfare by imposing a suitable tariff. To derive these results he developed a model (1907) in which nominal import and export prices were expressed as functions of the quantities of imports and exports respectively (with no cross-effects), each country being assumed to stabilize the value of its currency. The elasticities of demand for imports and supply of exports were defined as the reciprocals of the elasticities of these functions (with opposite sign). This has come to be known as the 'elasticity approach'. (For an interpretation of these demand and supply prices as prices relative to the price – assumed stabilized – of a non-tradable in a general-equilibrium model, see Chipman 1978.) Bickerdike derived formulas for the effect on national 'advantage' of a small tariff (p. 100n) and for the optimal tariff (p. 101n), and remarked – anticipating Lerner (1936) – that identical expressions would be obtained for an export tax. He noted that the optimal tariff depended only on the foreign elasticities (see also Kahn 1947); this apparent paradox was explained by Graaff (1949, p. 56). The now-familiar, simpler and more general optimal-tariff formula expressed in terms of Marshallian elasticity was first introduced by Johnson (1950), who showed its relation to Bickerdike's formula.

Edgeworth (1908, p. 544) showed that the positive sign of the denominator of Bickerdike's expression for the advantage from an incipient tariff followed from dynamic stability. A related stability condition was later derived by Bickerdike (1920) for the analysis of a regime of fluctuating exchange rates, and was obtained as a condition for a transfer to lower the paying country's exchange rate. Equivalent formulas were subsequently adopted by Robinson (1937, p. 194n) and Metzler (1948), and – for the special case indicated by Bickerdike of infinite elasticities of supply of exports – by Lerner (1944, p. 378).

Bickerdike's other contributions include two essays on local public finance (1902, 1912), a paper (1911) correcting a statement of Edgeworth's that price discrimination could improve upon competitive pricing, and papers on a number of other topics, the most noteworthy relating to business cycles and economic growth.

Although preceded by Carver (1903), Aftalion (1909, pp. 219–20) and Pigou (1912, pp. 144–5), Bickerdike (1914) may be considered one of the original developers of the acceleration principle (cf. Hansen 1927, p. 112; Haberler 1937, p. 87), providing a detailed numerical example and emphasizing (in contrast to Aftalion) the importance of durability of capital rather than the gestation period. Bickerdike regarded the phenomenon as an example of market failure. The paper was cited by Frisch (1931) – who erroneously attributed it to J.M. Clark – in the course of his criticism of Clark (1923) and reformulation according to which a deceleration of consumption will call forth a fall in gross investment only if it exceeds the rate of depreciation of capital. Bickerdike (1924, 1925) went on to develop an interesting mathematical model of economic growth according to which labour – the only factor – grows at a constant rate and produces only capital goods – of various durabilities and with various gestation periods – the services of which are consumed. On a path of balanced growth, the rate of interest is equal to the rate of growth, and interest is reinvested. The money supply grows at the same rate in order to maintain constant prices – or else it is constant and prices fall at a constant rate. Bickerdike's main object was to determine whether the process of saving benefited non-savers; in this he was not entirely successful, since his techniques limited him to balanced-growth paths. Nevertheless this work foreshadowed that of Lerner (1944, ch. 20) as well as many features of contemporary growth models, and attracted the attention of Hansen (1927, pp. 173ff).

Information on Bickerdike's life and work may be found in Jha (1963) and in Larson (1983, 1987), where other relevant literature is also cited. According to Larson, after Bickerdike's death his papers, including some 50 letters from Edgeworth and 20 from Edwin Cannan, passed into the hands of one Godfrey Alan Dick who died in Oxford in 1981. They are presumed lost.

See Also

- ▶ Elasticities Approach to the Balance of Payments
- ▶ Marshall–Lerner Condition

Selected Works

1902. Taxation of site values. *Economic Journal* 12: 472–484. Reprinted in Musgrave and Shoup (1959), 377–388.
1906. The theory of incipient taxes. *Economic Journal* 16: 529–535. Reprinted in Musgrave and Shoup (1959), 132–8.
1907. Review of protective and preferential import duties by A.C. Pigou. *Economic Journal* 17: 98–102.
1911. Monopoly and differential prices. *Economic Journal* 21: 139–148.
1912. The principle of land value taxation. *Economic Journal* 22: 1–15.
1914. A non-monetary cause of fluctuations in employment. *Economic Journal* 24: 357–370.
1920. The instability of foreign exchange. *Economic Journal* 30: 118–122.
1924. Individual and social interests in relation to saving. *Economic Journal* 34: 408–422.
1925. Saving and the monetary system, *Economic Journal* 35: 366–378.

Bibliography

Aftalion, A. 1909. La réalité des surproductions générales, 3rd instalment. *Revue d'économie politique* 23: 201–229.
Carver, T.N. 1903. A suggestion for a theory of industrial depressions. *Quarterly Journal of Economics* 17: 497–500.
Chipman, J.S. 1978. A reconsideration of the 'elasticity approach' to balance-of-payments adjustment problems. In *Breadth and depth in economics*, ed. J.S. Dreyer. Lexington: Heath.
Clark, J.M. 1923. *Studies in the economics of overhead costs*. Chicago: University of Chicago Press.
Clark, J.M. 1931. Capital production and consumer-taking – A reply. *Journal of Political Economy* 39: 814–816; A further word, 40: 691–693.
Edgeworth, F.Y. 1908. Appreciations of mathematical theories, III. *Economic Journal* 18: 392–403; 541–556. Reprinted as: Mr. Bickerdike's theory of incipient taxes and customs duties, In *Papers relating to political economy*, ed. F.Y. Edgeworth, vol. 2. London: Macmillan, 1925.
Frisch, R. 1931. The interrelation between capital production and consumer-taking. *Journal of Political Economy* 39: 646–654; A rejoinder, 40: 253–254; A final word, 40: 694.
de Graaff, J.V. 1949. On optimum tariff structures. *Review of Economic Studies* 17: 47–59.
Haberler, G. 1937. Prosperity and depression. 3rd ed. Lake Success: United Nations, 1946.
Hansen, A.H. 1927. *Business-cycle theory*. Boston: Ginn & Co..
Jha, N. 1963. *The age of marshall*. Patna: Novelty & Co; 2nd ed., London: Frank Cass, 1973.
Johnson, H.G. 1950. Optimum welfare and maximum revenue tariffs. *Review of Economic Studies* 19: 28–35.
Kahn, R.F. 1947. Tariffs and the terms of trade. *Review of Economic Studies* 15: 14–19.
Larson, B.D. 1983. The analysis of interests in the economics of Charles Frederick Bickerdike. Ph.D. dissertation, University of North Carolina, Chapel Hill.
Larson, B. 1987. Bickerdike's life and work. *History of Political Economy* 19: 1–22.
Lerner, A.P. 1936. The symmetry between import and export taxes. *Economica,* n.s. 3: 306–313.
Lerner, A.P. 1944. *The economics of control*. New York: Macmillan.
Metzler, L.A. 1948. The theory of international trade. In *A survey of contemporary economics*, ed. H.S. Ellis. Philadelphia: Blakiston.
Musgrave, R.A., and C.S. Shoup, eds. 1959. *Readings in the economics of taxation*. Homewood: Irwin.
Pigou, A.C. 1912. *Wealth and welfare*. London: Macmillan.
Robinson, J. 1937. The foreign exchanges. In *Essays in the theory of employment*, ed. Joan Dreyer. London: Macmillan; 2nd ed., Oxford: Basil Blackwell, 1947.

Bidding

Robert Wilson

Auctions are studied because they are market institutions of practical importance. Their simple procedural rules to resolve multilateral bargaining over the terms of trade enjoy enduring popularity. They also present simply several basic issues of price determination: the role of private information, the consequences of strategic behaviour, and the effect of many traders. These issues have

influenced the subject since the initial work of Vickrey (1961), the early contribution of Griesmer et al. (1967), and the influential dissertation by Ortega- Reichert (1968). Useful introductory surveys are by Engelbrecht-Wiggans (1980), Engelbrecht- Wiggans et al. (1983), Milgrom (1985), and MacAfee and McMillan (1986); bibliographies are in MacAfee and McMillan (1986), Stark and Rothkopf (1979), and Cassady (1967) provides an historical perspective.

This note supplements the entry on Auctions by summarizing some additional theoretical contributions to these issues. This literature relies on the game-theoretic perspective that emphasizes the implications of complete optimizing behaviour. Omitted here are the experimental studies that offer alternative predictions of bidder behaviour. It remains to determine which better describes the behavior of experienced, savvy bidders in the major auction markets. Although general equilibrium models of closed economies have been studied (Schmeidler 1980; Shapley and Shubik 1977; Wilson 1978), we focus on partial equilibrium models with bids and offers denominated in money terms. Also omitted are studies of markets with intermediaries such as brokers and specialists; models without private information (Dubey 1982; Milgrom 1986); and auctions in which losers also pay, as in price wars and wars of attrition.

In the traditional view, price determination is a consequence of market clearing: prices equate supply and demand. This clearing process is especially transparent in the case of auction markets. Essentially, auctions are markets with explicit trading rules that specify precisely how market clearing determines prices. For example, in a sealed-bid auction of one or more identical indivisible items, the (interval of) clearing prices is determined by intersecting the seller's supply schedule (reflecting the number of units available and announced reservation prices) with the demand schedule formed by arraying the buyers' bids in descending order. Non-discriminatory pricing sets the price at the highest rejected bid, discriminatory pricing charges each successful bidder the amount of his bid, and various intermediate cases are possible. Double auctions operate similarly except that the supply schedule is constructed by arraying the sellers' offers in ascending order. With divisible commodities, the aggregate schedules are obtained by constructing the sums of the traders' demand and supply schedules at each price. Oral auctions, such as the English auction, find a clearing price by calling for bids in ascending order. An oral double auction, or 'bid–ask' market, allows free outcry of bids and offers that can be accepted immediately and therefore depends on participants' judgments about the likely clearing price.

The variety of possible procedural rules is large, so theoretical studies emphasize the characterization of efficient trading rules, such as rules that are optimal for the buyers or the sellers. The design of trading rules is subject to the incentive compatibility constraints induced by the traders' private information and the option of any trader to forego participation or trade. Auctions are especially restrictive trading mechanisms because their rules are specified independently of information about the distribution of traders' attributes, even if this information is common knowledge. On the other hand, auctions have been important market institutions for millennia precisely because they are efficient or nearly so in a wide variety of environments.

Much of the theory of efficient trading rules studies 'direct revelation' games in which, in equilibrium, each trader's action consists of a direct report of his private information. This approach loses no generality in static models but the resulting optimal rules depend on the distribution of traders' attributes: only in special cases can they be implemented fully as auctions. (In the extreme case of highly correlated private information, an *optimal* trading rule can be designed by the seller to extract most or all of the potential revenue (Crémer and McLean 1985; Myerson 1981).) The theory therefore divides between the study of auctions, in which traders' strategies take account of the distribution of attributes, and the study of optimal direct revelation games, in which the trading rule incorporates this data. We concentrate on auctions here, but mention intersections with the general theory.

A trading rule specifies each trader's feasible actions and the prices and trades resulting from their joint actions. Models also specify each trader's information and preferences. Typically each trader i knows privately an observation s_i affecting his preferences, and the restrictive assumption is adopted that the joint probability distribution of these observations and any salient unobserved random variable v is common knowledge among the participants. (The observation s_i is often taken to be real valued for simplicity; it could be the bidder's posterior certainty-equivalent valuation of the item based on his private information.) A strategy therefore specifies a trader's actions depending on his observation and any further observations (such as others' bids) made in process. Trader i's expected utility u_i depends on the received quantity q_i, the price (s) p_i at which these units are traded, his observation s_i, the array $S_i = \{s_j \mid j \neq i\}$ of others' observations, and possibly on other variables v.

Interesting special cases of the probabilistic structure are: independent and identically distributed (iid) observations; conditionally iid observations given v; and more generally, affiliated observations (e.g., nonnegative correlation on any rectangle). In each case assume that the (conditional) distribution of an observation satisfies the monotone hazard rate or likelihood ratio property. Most of the familiar probability distributions satisfy these assumptions; e.g., log-normal distributions are often used in applications to oil-lease bidding.

Interesting special cases of the preference structure for a single item include: private values, $u_i = s_i - p_i$; a common value, $u_i = v - p_i$; mixed values, $u_i = u(s_i, v) - p_i$, where u is increasing; and private-value cases with common risk aversion, $u_i = U(s_i - p)$, where U is increasing and concave. Relevant features are summarized in the expected utility $u(s_i, S_i) = E\{\bar{u}; (s_i, v) | s_i, S\}$.

Other features are also addressed in some formulations: the seller's optimal reservation price, a trader's option to obtain costly further observations to improve his information, bids submitted jointly by syndicates of traders, and entry fees and auxilliary contingent payments such as royalties. (Bidding on the royalty rather than the price has been used in auctions of oil leases.) Uncertainty about the number of bidders is easily included if this number is independent of the bidders' observations; however, somewhat different comparative statics results ensue. If there are bid preparation costs, exposure constraints (total amount of bids submitted) or portfolio motives, then participation in an auction is itself a strategic action and may involve randomization if there are too many potential bidders for all to expect to recoup their costs. If information is costly and subject to choice then even with many bidders there is typically an upper bound on the bidders' total expenditures and each bidder may choose to collect relatively little information (Matthews 1984). Repeated auctions introduce novel features, such as reputation effects, that severely alter the results; e.g., one bidder with privileged information can win systematically (Bikhchandani 1985).

Most theoretical studies assume that the traders' strategies form a Nash equilibrium, or in dynamic formulations, a sequential equilibrium: each strategy in each contingency is optimal for the remainder of the game. For many auction models the equilibrium strategies can be characterized elegantly in terms of the joint distribution of observations and bids (Milgrom and Weber 1985). If the bidders (on the same side of the market) are positioned symmetrically *ex ante* then one focuses on the symmetric equilibrium in which all bidders use the same strategy, which is an increasing function of one's observation. A large class of symmetric discriminatory auctions have *only* symmetric equilibria (Maskin and Riley 1986); they are usually characterized by differential equations, as illustrated for various cases in Milgrom and Weber (1982), Reece (1978), and Wilson (1977, 1985). In non-discriminatory auctions a single equation specifies the optimal bid as the most one would be willing to pay conditional on one's observation being the most optimistic. Results about symmetric equilibria are fairly robust: examples indicate that under- or over-bidding by one participant engenders a similar but muted response by others, and the difference from the symmetric equilibrium varies smoothly. In sealed-bid

discriminatory auctions with iid private values, one's bid is essentially the conditional expectation of the highest rejected valuation given that one's valuation is acceptable. An analogous property applies to mixed-value preferences. The important asymmetric cases occur when some bidders' information is superior to others' (e.g., direct information about v); in these cases any bidder with strictly inferior information obtains expected profits at most zero, and bidders may use randomized strategies. If all bidders have private information (with a positive density satisfying technical restrictions) then typically equilibrium strategies are not randomized and positive expected profits result.

We summarize results mainly for the special probabilistic and preference structures mentioned above, and for symmetric equilibria. Also, multiple-item auctions introduce few novelties when there is a single seller offering a fixed supply of identical items and each bidder wants at most one, so we focus on the single item case. An exception is a 'share auction', in which bidders offer demand schedules for shares of a divisible item in fixed supply: in this case there can be a continuum of symmetric equilibria, and the seller's expected revenue can be unaffected by more bidders (Wilson 1979).

A main effect of risk aversion is to increase bids in symmetric discriminatory auctions with iid private values. The seller can enhance this effect by imposing an entry fee (preferably decreasing in the amount of the bid and ultimately negative for the highest bids) (Matthews 1983; Maskin and Riley 1984). Risk aversion induces bidders to bid higher under discriminatory pricing, and in fact this rule makes the winning bid a less risky random variable. The seller therefore prefers discriminatory pricing, and more so if he too is risk averse. However, if bidders have decreasing absolute risk aversion (ARA), they have the reverse preference (Matthews 1987). With constant (or zero) ARA, a bidder's higher price with discriminatory pricing is exactly balanced by the riskier price associated with nondiscriminatory pricing. With affiliated observations, the bidders prefer discriminatory pricing if they have constant ARA, and will be indifferent again at some degree of decreasing ARA. Affiliation biases the seller's preferences in the opposite direction, towards nondiscriminatory pricing.

Hereafter we assume no risk aversion. Then, in the iid private-values model of bidders' preferences, the seller's expected revenue is the same for discriminatory and non-discriminatory pricing (Harris and Raviv 1981a, b; Myerson 1981; Riley and Samuelson 1981). Moreover, subject to a technical restriction, either of these is optimal among all possible trading rules provided the seller adopts an optimal reservation price (Harris and Raviv 1981a, b; Myerson 1981). With more general preferences, whenever $\bar{u}$ is increasing affiliation produces a distinct preference of the seller for (and the bidders against) non-discriminatory pricing vs. discriminatory; indeed, the seller further prefers an oral auction (Milgrom and Weber 1982). This illustrates the 'linkage principle': the seller wants to reduce the bidders' profits from their private information, and auction rules that reveal affiliated information publicly (inferences from bids in the case of oral auctions) or otherwise positively link one bidder's price to another's bid (non-discriminatory pricing) are advantageous when observations are affiliated and therefore positively correlated. Similarly, the seller prefers to reveal publicly any relevant affiliated information he has so as to reduce the bidders' informational advantages *vis-à-vis* each other. (However, revealing non-affiliated information may be disadvantageous, and in particular this applies to the number of bidders, even when it is independent of other data (Matthews 1987).) The seller can gain further by conditioning payments *ex post* on realized values, as in the case of a royalty (Riley 1986).

The main results about bidders' strategies in single-item sealed-bid discriminatory auctions can be summarized for bidders with symmetric conditionally iid mixed-value preferences. *Ex ante* each bidder has an equal chance of winning and the bidder with the most optimistic observation is predicted to win, namely i wins in the event $W(s_i) = \{s_i, > \max S_i|s_i,\}$. (Failure to recognize that winning is an informative event, signalling that others' observations were less optimistic, is called the winner's curse (Capen et al. 1971); it is

Bidding, Table 1 Examples of equilibrium bidding strategies lognormal distributions

$u_i = s_{i1}v$, $\sigma 2_1 + \sigma^2{}_2 = 1.0$, $\sigma^2{}_2 = 0.36$					
Number of bidders	(n)	2	4	8	16
Bid factor	(%)	30.9	39.0	40.5	39.6
Expected profit	(%)	53.45	33.85	23.78	17.82

distressingly common in practice as well as in experiments. The implications of the fact that the maximum of several unbiased estimates is biased upward are apparently difficult to appreciate.) The most that i can profitably bid is therefore $\widehat{u}(s_i) = E\{\overline{u}(s_i, S_i)|W(s_i)\}$, whereas the optimal bid is less than this, by a percentage that is of the order of $1/n$ when there are n bidders, reflecting the bidder's monopoly rent both in terms of the limited number of bidders and the advantage of his private information, which are the two sources of bidders' expected profits. (In a nondiscriminatory auction, i bids $\hat{u}(s_i)$ computed from $\widehat{W}(s_i) \equiv \{s_i \max S_i | s_i\}$ but in equilibrium pays $\hat{u}(\max S_i)$ if he wins.) With many bidders these rents are dissipated and, remarkably, the winning bid conveys essentially all the information about v contained in $\max_i s_i$,. In the common value model, the winning bid is a consistent estimator of the value whenever *any* consistent estimator exists that is a function of $\max_i s_i$,: the winning bid is asymptotically as good an estimator as is possible from extrema of the bidders' observations. In particular, if the relative likelihood of a large observation is small for smaller values of v, then the maximum bid converges in probability to v as the number of bidders increases (Milgrom 1979a, b; Palfrey 1985; Wilson 1977).

These features are reflected in the detailed calculations reported for models of oil-lease bidding (Reece 1978). Other examples are shown in Table 1, which exhibits the equilibrium strategies for a model that roughly approximates firms' bidding for oil leases. Each bidder, observes $s_{\mathrm{i}} = (s_{i1}, s_{i2})$ and $u(s_i, v) = s_{i1}, v$, where s_{i1} represents a private factor (e.g., price or discount factor), and s_{i2} represents an estimate of the common factor v. Assume that, conditional on a location parameter s, the private factors are conditionally independent and ln s_{i1} has mean ln $\overline{s}_1$; and variance of; and marginally ln $\overline{s}_1$ has variance $\overline{\sigma}_1^2$. Similarly, conditional on v the estimates are conditionally independent and ln s_{i2} has mean ln v and variance $\overline{\sigma}_2^2$; and marginally ln v has variance $\overline{\sigma}_2^2$. Consider the case adapted to the empirical fact that for Gulf of Mexico oil leases the logarithm of the bids typically has conditional variance about 1.0 whereas the estimating precision implies a variance of about 0.36 given that the prior variances $(\overline{\sigma}_1^2, \overline{\sigma}_2^2)$ are comparatively so large that they can be considered infinite: assume that the conditional variance of the private factors accounts for the difference. In this case, the symmetric equilibrium bidding strategy specifies that each firm submits a bid that is a specified fraction (the bid factor) of the product of its private factor and its posterior expectation of the common factor given its estimate. The tabulation shows the percentage bid factor for four numbers n of bidders, assuming the seller's reservation price is zero, and it shows the winning bidder's expected percentage profit. The seemingly low bid factors are necessary to avoid the winner's curse; whereas the surprisingly large profit percentages reflect the role of the private valuation factors.

Analogous models in which a bidder can increase the precision of his information at increasing cost differ in that, even though bidders' total expenditures converge to a positive level as the number of bidders increases, each bidder's expenditure converges to zero. In this case the winning bid is not a consistent estimator of the common value v and the seller's expected revenue is reduced by the amount of the bidders' total expenditure on information, since in equilibrium this is necessarily recouped in expectation by the participating bidders (Matthews 1984). An important policy conclusion is that bidders' expenditures on information are inefficiently large.

Single-item auctions with dynamic rules add a few new aspects. In a Dutch auction an exogenously specified price is lowered until a bidder

accepts. This rule induces a game that for the bidders is strategically equivalent to a sealed-bid auction; it is also payoff equivalent unless they are impatient to trade, in which case a bidder is concerned about the sum of the interest rate and the hazard rate that trade will be unsurped by a competitor. For the seller it differs if he cannot commit to forego trading by stopping the price at a reservation price exceeding his valuation. In the iid private-values model of preferences, a generalized multi-item Dutch auction is an optimal selling strategy for a monopolist seller whenever potential demand exceeds supply (Harris and Raviv 1981a, b). Auctions with exogenously ascending prices, in which the items are awarded to the remaining bidders at the price at which the last of the others drops out, are a form of nondiscriminatory auction; but with affiliated observations, a bidder's strategy accounts for the learning enabled by seeing the prices at which others drop out – an instance of the linkage principle.

Double auctions have been studied only for the case of iid private values and nondiscriminatory pricing. The price chosen is the midpoint of the interval of clearing prices derived from intersecting the schedules of bids (arrayed in descending order) and offers (arrayed in ascending order). Such an auction is actually an *ex ante* efficient trading rule for the case of one buyer and one seller with values distributed uniformly on the same interval (Chatterjee and Samuelson 1983; Myerson and Satterthwaite 1983); and by implication from the previous results for auctions, for one buyer *or* one seller. With several buyers and sellers and fairly general distributions, the *ex ante* efficient trading rule bears a strong resemblance to a double auction and has the remarkable property that the expected efficiency losses (compared to *ex post* efficiency) from strategic behaviour decline nearly quadratically to zero as the numbers of traders increase (Gresik and Satterthwaite 1984). The weaker criterion of *interim* efficiency requires that no other trading rule is sure to improve every trader's expected gains from trade: a double auction satisfies this criterion if there are sufficiently many buyers and sellers (Wilson 1985).

Oral multi-item discriminatory double auctions, allowing free outcry of bids and offers, are the most important practically (e.g. commodity markets) and the most challenging theoretically. Since trades are consummated in process at differing prices, 'market clearing' is dynamic and, for example, traders with extra-marginal valuations in the static sense can obtain gains from trade early on. Since traders are continually motivated to estimate the distribution of subsequent bids and offers, the learning process is a key feature. Theoretical studies have been attempted for both complete equilibrium models (Wilson 1987) and others invoking some plausible behavioral assumptions (Easley and Ledyard 1982; Friedman 1984). These studies aim to explain the dramatic efficiency attained in experiments and the tendency for transaction prices to approximate or converge to the static Walrasian clearing prices (Smith 1982; Plott and Sunder 1982), especially with replication even when the subjects lack a base of common knowledge about distributional features. The efficiency realized in experimental settings is a major puzzle deserving better theoretical explanations.

In summary, auctions are important market institutions that ensure market clearing via explicit trading rules that are independent of the distribution of preferences and information among the participants. Over wide ranges of models of preferences and information, these trading rules are *ex ante* or *interim* efficient or nearly so, and both practically and experimentally they are evidently robust. The theory elaborates these properties and demonstrates the role of private information and strategic behaviour. The explicit construction of equilibrium strategies establishes the magnitudes of these effects and enables comparisons of trading rules, preference structures, informational conditions, and the number of participants; and additionally it explains phenomena such as the winner's curse that stem from adverse selection effects when there is dispersed information. Some models predict that the choice of pricing rule is inconsequential because bidders alter their strategies to compensate: the market clearing condition is the main determinant of welfare consequences.

In relation to the general economic theory of markets the theory of auctions addresses the special case of markets with explicit market-clearing trading rules and elaborates in fine detail the determination of prices and the efficiency and distributional consequences of particular assumptions about the attributes of participants. This endeavour is a useful step in the construction of a general theory of the microstructure of markets that encompasses the full range from bilateral bargaining to 'perfectly competitive' markets.

See Also

► Auctions

Bibliography

Bikhchandani, S. 1985. *Reputation in repeated second price auctions*. Research Paper 815, Stanford Business School, July 1985; Also in 'Market games with few traders', PhD dissertation, 1986.

Capen, E., R. Clapp, and W. Campbell. 1971. Competitive bidding in high-risk situations. *Journal of Petroleum Technology* 23: 641–653.

Cassady Jr., R. 1967. *Auctions and auctioneering*. Berkeley: University of California Press.

Chatterjee, K., and W. Samuelson. 1983. Bargaining under incomplete information. *Operations Research* 31: 835–851.

Crémer, J., and R. McLean. 1985. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica* 53: 345–363.

DeBrock, L., and J. Smith. 1983. Joint bidding, information pooling, and the performance of petroleum lease auctions. *Bell Journal of Economics* 14: 395–404.

Dubey, P. 1982. Price–quantity strategic market games. *Econometrica* 50: 111–126.

Easley, D., and J. Ledyard. 1982. *A theory of price formation and exchange in oral auctions*. Discussion Paper 461, Northwestern University, 1982.

Engelbrecht-Wiggans, R. 1980. Auctions and bidding models: A survey. *Management Science* 26: 119–142.

Engelbrecht-Wiggans, R., P.R. Milgrom, and R.J. Weber. 1983a. Competitive bidding and proprietary information. *Journal of Mathematical Economics* 11: 161–169.

Engelbrecht-Wiggans, R., M. Shubik, and R. Stark (eds.). 1983b. *Auctions, bidding, and contracting: Uses and theory*. New York: New York University Press.

Friedman, D. 1984. On the efficiency of double auction markets. *American Economic Review* 74: 60–72.

Gresik, T., and M. Satterthwaite. 1984. *The rate at which a simple market becomes efficient as the number of traders increases: An asymptotic result for optimal trading mechanisms*. Discussion Paper 641, Northwestern University, 1984.

Griesmer, J., R. Levitan, and M. Shubik. 1967. Towards a study of bidding processes, part four: Unknown competitive costs. *Naval Research Logistics Quarterly* 14: 415–433.

Harris, M., and A. Raviv. 1981a. A theory of monopoly pricing schemes with demand uncertainty. *American Economic Review* 71: 347–365.

Harris, M., and A. Raviv. 1981b. Allocation mechanisms and the design of auctions. *Econometrica* 49: 1477–1499.

Harris, M., and R. Townsend. 1981. Resource allocation under asymmetric information. *Econometrica* 49: 33–64.

Holt Jr., C.A. 1979. Uncertainty and the bidding for incentive contracts. *American Economic Review* 69: 697–705.

Holt Jr., C.A. 1980. Competitive bidding for contracts under alternative auction procedures. *Journal of Political Economy* 88: 433–445.

MacAfee, E.P., and J. McMillan. 1987. Auctions and bidding. *Journal of Economic Literature* 25: 699–738.

Maskin, E., and J. Riley. 1984a. Monopoly with incomplete information. *Rand Journal of Economics* 15: 171–196.

Maskin, E., and J. Riley. 1984b. Optimal auctions with risk averse buyers. *Econometrica* 52: 1473–1518.

Maskin, E., and J. Riley. 1986. *Existence and uniqueness of equilibrium in sealed high bid auctions*. Discussion Paper 407, University of California at Los Angeles, Mar 1986.

Matthews, S. 1983. Selling to risk averse buyers with unobservable tastes. *Journal of Economic Theory* 30: 370–400.

Matthews, S. 1984a. Information acquisition in discriminatory auctions. In *Bayesian models in economic theory*, ed. M. Boyer and R. Kihlstrom. Amsterdam: North-Holland.

Matthews, S. 1984b. On the implementability of reduced form auctions. *Econometrica* 52: 1519–1522.

Matthews, S. 1987. Comparing auctions for risk averse buyers: A buyer's point of view. *Econometrica* 55: 633–646.

Milgrom, P.R. 1979a. *The structure of information in competitive bidding*. New York: Garland Publishing Co.

Milgrom, P.R. 1979b. A convergence theorem for competitive bidding with differential information. *Econometrica* 47: 679–688.

Milgrom, P.R. 1981. Rational expectations, information acquisition, and competitive bidding. *Econometrica* 49: 921–943.

Milgrom, P.R. 1985. The economics of competitive bidding: A selective survey. In *Social goals and social*

organization, ed. L. Hurwicz, D. Schmeidler, and H. Sonnenschein. Cambridge: Cambridge University Press.
Milgrom, P.R. 1987. Auction theory. In *Advances in economic theory 1985*, ed. T. Bewley. Cambridge: Cambridge University Press.
Milgrom, P.R., and R.J. Weber. 1982a. A theory of auctions and competitive bidding. *Econometrica* 50: 1089–1122.
Milgrom, P.R., and R.J. Weber. 1982b. The value of information in a sealed-bid auction. *Journal of Mathematical Economics* 10: 105–114.
Milgrom, P.R., and R.J. Weber. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10: 619–632.
Moore, J. 1984. Global incentive constraints in auction design. *Econometrica* 52: 1523–1525.
Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operations Research* 6: 58–73.
Myerson, R.B., and M.A. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29: 265–281.
Ortega-Reichert, A. 1968. *Models for competitive bidding under uncertainty.* Technical Report 8, Operations Research Department Stanford University.
Palfrey, T.R. 1985. Uncertainty resolution, private information aggregation and the Cournot competitive limit. *Review of Economic Studies* 52: 69–83.
Plott, C.R., and S. Sunder. 1982. Efficiency of experimental security markets with insider information: An application of rational expectations models. *Journal of Political Economy* 90: 663–698.
Reece, D.K. 1978. Competitive bidding for offshore petroleum leases. *Bell Journal of Economics* 9: 369–384.
Reece, D.K. 1979. Alternative bidding mechanisms for offshore petroleum leases. *Bell Journal of Economics* 10: 659–669.
Riley, J. 1986. *Ex post information in auctions*. Discussion Paper 367, University of California at Los Angeles, Mar.
Riley, J., and W. Samuelson. 1981. Optimal auctions. *American Economic Review* 71: 381–392.
Schmeidler. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48: 1585–1593.
Shapley, L., and M. Shubik. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85: 937–968.
Smith, V. 1982. Microeconomic systems as experimental science. *American Economic Review* 72: 923–955.
Stark, R.M., and M.H. Rothkopf. 1979. Competitive bidding: A comprehensive bibliography. *Operations Research* 27: 364–390.
Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16: 8–37.
Vickrey, W. 1962. Auctions and bidding games. In *Recent advances in game theory*, ed. O. Morgenstern and A. Tucker. Princeton: Princeton University Press.
Wilson, R. 1977. A bidding model of 'perfect' competition. *Review of Economic Studies* 44: 511–518.
Wilson, R. 1978. Competitive exchange. *Econometrica* 46: 557–585.
Wilson, R. 1979. Auctions of shares. *Quarterly Journal of Economics* 93: 675–689.
Wilson, R. 1985. Incentive efficiency of double auctions. *Econometrica* 53: 1101–1116.
Wilson, R. 1987. Equilibria of bid-ask markets. In *Arrow and the ascent of economic theory: Essays in honour of Kenneth J. Arrow*, ed. G. Feiwel. London: Macmillan.

Bidding Rings

John Asker

Abstract

A bidding ring is a collection of bidders who collude in an auction in order to gain greater surplus by depressing competition. This entry describes some typical bidding rings and provides an introduction to the related theoretical and empirical literature.

Keywords

Cartel; Antitrust; Bidding ring; Bid rigging; Sherman Act; Auctions; Price fixing; Bid rotation; Collusion

JEL Classifications

D44; K21; L41; L12

When bidders in an auction collude in order to diminish competition between themselves, and hence earn greater surplus, the resulting cartel is often referred to as a bidding ring. The act of colluding in an auction is often referred to as 'bid rigging'. Bidding rings are illegal in most jurisdictions. In the USA, for example, a bidding ring is a violation of the Sherman Act and is punished by fines for both individuals and firms, and by jail time for those individuals involved.

Canonical examples of bidding rings include the 'Electrical Conspiracy' in the 1950s, in which 29 suppliers of industrial electrical generators and equipment colluded in first price sealed bid procurement auctions (Smith 1961; McAfee and

McMillian 1992). This ring used a bid rotation scheme in which each ring member was allocated a phase of the Moon. The phase of the Moon at the time of the auction determined which of the ring members had the right to bid, free from competition from other members of the ring. Another example, this time in an ascending price auction, and involving an explicit sidepayment system, was the ring adopted by 81 book dealers in the auction of the library of Ruxley Lodge in 1919 (Freeman and Freeman 1990) and Porter (1992)). After buying up the contents of the library free from internal competition, the ring members met in a sequence of knockout auctions which reallocated the contents of the library to those ring members who valued them the most. (A knockout auction is an auction conducted among ring members.) Participation in the knockouts became more restricted as the sequence progressed. The proceeds of each knockout were shared equally among participants, thus generating a system of sidepayments that increased with the participation (and presumably importance) of each ring member (Graham et al. (1990) describe similar cartels).

Importantly, many examples exist of bidding rings with many members, providing counterexamples to the common presumption that collusion is prohibitively difficult in markets with large numbers of participants.

The theoretical literature on bidding rings tends to focus on how the ring can allocate bids and transfers to its members in a way that is incentive compatible and extracts the greatest surplus for the cartel, given a series of institutional features. These institutional features include: the format of the auction; whether explicit sidepayments are feasible; the interdependence of bidders' private information (e.g. common values (CV) vs. independent private values (IPV)); the extent to which ring members are *ex ante* symmetric; the extent to which the mechanism should be budget balancing (i.e. whether within-ring transfers net to zero); and whether the ring faces competition from outside bidders. All of these features can affect the form of mechanism used by the ring to coordinate bidding and allocate surplus. The enforcement of the obligations arising from the ring mechanism is most often attributed to repeated game strategies (Athey and Bagwell 2001).

In an IPV environment the central challenge facing the ring is getting each ring member to reveal their valuation for the object at auction. The problem is that the bids and within-ring transfers will often depend on the valuations ring members report to the ring. This potentially gives ring members incentives to misreport their valuations in the hope of gaining a greater share of the collusive surplus.

In IPV first price sealed bid auctions, McAfee and McMillian (1992) show that without explicit sidepayments the best an (all-inclusive) ring can do is to randomize over which ring member wins and for every ring member to merely bid the reserve price. As they point out, this closely resembles the phases of the Moon scheme used in the 'Electrical Conspiracy'. Such a scheme must lead to inefficient allocations and hence diminishes social welfare. Where sidepayments are feasible and ring members are *ex ante* symmetric the optimal ring mechanism can be implemented using a first price sealed bid knockout auction prior to the auction. The winner of this knockout gets the right to bid in the auction and the revenue raised in the knockout is shared equally among the ring members.

Knockout auctions also feature centrally in the theory of collusion in IPV ascending price auctions (see Mailath and Zemsky (1991) for characterization results). Graham et al. (1990) depart from the standard mechanism design approach, investigating the use of the Shapely value to allocate sidepayments to ring bidders. Despite the fact that such a payment scheme can lead to somewhat perverse bidding incentives, the scheme they describe mirrors both the Ruxley Lodge example above and the ring described in Asker (2009).

In common value settings the central issue is information aggregation. Hendricks et al. (2008) point out that the ring can increase aggregate surplus by providing a way to aggregate bidders' signals of the underlying value of the object. However, some bidders may prefer a non-cooperative auction, as the ring's

sidepayment scheme can lead bidders with lower signals to benefit at the expense of bidders with higher signals. They provide conditions under which *ex post* efficiency, budget balance and individual rationality are incompatible elements of an indirect mechanism.

Empirical work on bidding rings suffers from the difficulty of getting highquality data on what is an illegal, and hence secretive, activity. The majority of empirical papers on bidding rings consider the statistical detection of bidding patterns consistent with cartel activity. A smaller body of work examines how bidding rings are structured and the extent to which they appear to distort market outcomes.

The statistical detection of bidding rings proceeds by writing down a model of the suspected ring and then comparing the observed bidding pattern to that of the modelled ring and a non-collusive benchmark. For instance, Porter and Zona (1993) examine bidding in highway paving contracts on Long Island, comparing the rank distribution of bids submitted by (known) ring and non-ring bidders. They find the order of the less competitive ring bids is not explained by capacity utilization rates, whereas the order of less competitive non-ring bids is explained by the respective firms' capacity utilization rates. They interpret this as being consistent with the operation of the bidding ring. Bajari and Ye (2003) propose a similar detection scheme.

The few papers that have studied the structure of known cartels and their impact on market outcomes have found that cartels come surprisingly close to implementing optimal mechanisms. Pesendorfer (2000) examines bidding rings in first price sealed bid auctions for contracts to supply school milk in Florida and Texas using data collected during the prosecution of the rings. The Florida ring used a market division scheme while the Texas ring used an system of explicit side-payments. Pesendorfer draws inferences about the underlying structure of the rings from observed bidding data and concludes that the rings were using mechanisms that were close to optimal. Asker (2010) also uses data collected during a prosecution, this time from a ring operating in ascending price auctions for collectable stamps. Asker concludes that the ring captures 72% of the surplus generated by the theoretically optimal ring and, interestingly, imposes damages on both sellers and competing bidders (by pushing prices above competitive levels at times and also by introducing inefficient allocations).

See Also

- ▶ Anti-trust Enforcement
- ▶ Auctions (Theory)
- ▶ Cartels
- ▶ Collusion

Bibliography

Asker, J. 2010. A study of the internal organization of a bidding cartel. *American Economic Review* 100(3): 724–762.

Athey, S., and K. Bagwell. 2001. Optimal collusion with private information. *RAND Journal of Economics* 32: 428–465.

Bajari, P., and L. Ye. 2003. Deciding between competition and collusion. *Review of Economics and Statistics* 85: 971–989.

Freeman, A., and J. Freeman. 1990. *Anatomy of an auction: Rare books at Ruxley Lodge 1919*. London: The Book Collector.

Graham, D., R. Marshall, and J.-F. Richard. 1990. Differential payments within a bidder coalition and the Shapley value. *American Economic Review* 80(3): 493–510.

Hendricks, K., R. Porter, and G. Tan. 2008. Bidding rings and the winner's curse. *RAND Journal of Economics* 39: 1018–1041.

Mailath, G., and P. Zemsky. 1991. Collusion in second price auctions with heterogenous bidders. *Games and Economic Behaviour* 3: 467–486.

McAfee, R.P., and J. McMillian. 1992. Bidding rings. *American Economic Review* 82(3): 579–599.

Pesendorfer, M. 2000. A study of collusion in first price auctions. *Review of Economic Studies* 67: 318–411.

Porter, R., and D. Zona. 1993. Detection of bid rigging in procurement auctions. *Journal of Politi cal Economy* 101: 518–538.

Porter, R.1992. Review of: Freeman, A. and Freeman, J. (1990) Anatomy of an auction: Rare books at Ruxley Lodge 1919. *Journal of Political Economy* 100(2): 433–436

Smith, R. 1961. The incredible electrical conspiracy. *Fortune* 63: 132–180. 161–224.

Bilateral Monopoly

James W. Friedman

A bilateral monopoly is a market that is characterized by one firm or individual, a monopolist, on the supply side and one firm or individual, a monopsonist, on the demand side. The input markets of the monopolist and the output market of the monopsonist can be of any form. The essential ingredient is the single seller–single buyer situation. Because a buyer and a seller of a product, perforce, do business with each other, they are clearly able to make legally binding agreements. This contrasts with firms in the same industry, which do not sell to one another, and which are often precluded by anti-collusion laws from making legally enforceable contracts. Of course, it is also possible to view bilateral monopoly noncooperatively.

The following coverage is chronological, starting first with the cooperative treatment due to Edgeworth (1881) and Marshall (1890). The noncooperative formulations, due to Wicksell (1925) and Bowley (1928) are considered next, along with Bowley's reformulation of the Marshallian cooperative contribution. Finally, bilateral monopoly is viewed in game-theoretic terms as a two-player cooperative game, principally in the manner of Nash (1950).

Bilateral monopoly is a special instance of two-person trade; therefore, the natural starting point is Edgeworth's (1881, pp. 20–30) well known analysis. Suppose the two agents, A and B, have utility functions $u^{\mathrm{A}}(x, y)$ and $u^{\mathrm{B}}(X - x, X - x, Y - y)$, where x and y are quantities of two goods consumed by A. The totals available to the pair are X and Y, respectively; hence, the consumption of B is $(X - x, Y - y)$. Edgeworth proposed that the two persons would trade to a Pareto optimal outcome that left each at least as well off as he would be in the absence of trade.

Marshall (1890, Appendix F and Mathematical Note XII) noted that, if both persons' marginal utilities are constant for one of the goods (say y), then any Pareto optimal trade will involve a fixed quantity of the other good (x). This is easily seen by recalling that a Pareto optimal (interior) trade requires equality of the two traders' marginal rates of substitution,

$$\frac{u_x^A(x,y)}{u_y^A(x,y)} = \frac{u_x^B(X-x,Y-y)}{u_y^B(X-x,Y-y)}$$

and then invoking Marshall's condition, which is and

$$u^A(x,y) = v^A(x) + ay$$

and

$$u^B(X-x,Y-y) = v^B(x-x) + b(Y-y).$$

Note that Marshall's condition is actually that the two traders each have utility functions that are separable and linear with respect to one good. Equality of the marginal rates of substitution is given by

$$\frac{v_x^A(x)}{a} = \frac{v_x^B(X-x)}{b}$$

which is independent of y.

Bowley (1928) put Marshall's result in the following standard bilateral monopoly model: suppose the seller has the profit function $\pi^{\mathrm{A}} = rx - C(x)$, where r is the firm's selling price, x is the amount sold, and $C(x)$ is the firm's total cost function; the buyer's profit function is $\pi^{\mathrm{B}} = f(x) - rx$, where x is the buyer's only input, and $f(x)$ is its total revenue as a function of the sole input x. (That is, $f(x) = d[h(x)] \cdot h(x)$, where h is the production function and d is the inverse demand function for the firm.) The decision variables are x and r, and the Pareto optimality condition is

$$\frac{\partial \pi^A/\partial x}{\partial \pi^A/\partial r} = \frac{\partial \pi^B/\partial x}{\partial \pi^B/\partial r} \quad \text{or} \quad \frac{r - C'(x)}{x} = \frac{f'(x) - r}{-x}$$

The latter condition is independent of r and is equivalent to $f'(x) = C'(x)$, which states that the

marginal revenue of the buying firm should equal the marginal cost of the selling firm – the condition for joint profit maximization. The way that the joint profit is split depends upon r, the transfer price between the two firms.

An equilibrium in a noncooperative vein was suggested by Wicksell (1925, pp. 223–5) and developed by Bowley (1928). Wicksell's equilibrium features a price announcement by the seller, followed by a quantity selection by the buyer. The seller is committed to deliver whatever amount the buyer wishes at the named price. Thus, the seller is a Stackelberg (1934) leader that knows the buyer will maximize $f(x) - rx$ with respect to x, and with r assumed constant. This allows the seller to solve $f'(x) = r$ for x as a function of r [denote this $x = \phi(r)$] and use this in its own profit function, $r\phi(r) - C[\phi(r)]$, which it then maximizes with respect to r to find the best price to announce.

In addition to working out the details of the foregoing model, Bowley suggested an alternative in which the roles of the two firms are exactly reversed: the buyer announces a price at which it will buy any quantity the seller cares to deliver, and the seller then chooses an amount to transact. The buyer can calculate the optimal choice of x for the seller as a function of r and then use this information to determine its most profitable price.

These noncooperative outcomes are not in general Pareto optimal; therefore, they are implausible in a setting such as this where there are only two agents who can discuss a transaction with one another and who are quite able to make binding agreements that do give them Pareto optimal outcomes.

Another way to visualize the possible outcomes in a bilateral monopoly is in the profit (or utility) space of the agents. In the Marshall–Bowley model the payoff possibility frontier is a straight line of slope $-b/a$. Were the two players a firm and a labour union, then, depending on the utility function assigned to the union, the payoff possibility frontier need not be a straight line. The union's utility function might well depend on the wage rate, the number of workers employed, and the average hours worked per employee. The representation of the model in profit, or utility, space is useful in approaching the model as a game-theoretic bargaining problem.

Perhaps the most famous two person cooperative game solution is that due to Nash (1950), in which there is a *threat outcome* that would prevail in the absence of agreement between the two players, with the bargained outcome being on the payoff possibility frontier at that point where the product of the players' gains from cooperation is maximized. Though this product maximization rule seems arbitrary on the surface, it is implied by several axioms that are plausible. For the bilateral monopoly model the threat of each firm is to refuse to trade with the other. This threat would force zero profit onto the buyer and $-C(0)$ onto the seller.

Nash's approach can be enriched in several ways. First, the threat of no trade need not leave the firms at profits of 0 and $-C(0)$. Perhaps the seller could enter the buyer's line of business. Similarly, the buyer may have other options open: there may be substitutes for the input supplied by the seller that are more expensive or less effective. If both possibilities hold at once, then 'no trade' does not completely specify the threat situation. The firms could become duopolists in a vertically integrated industry, and carry out threats in terms of the output levels that they decide to produce. This leads to a *variable threat* game, analysed by Nash (1953).

Neither of the Nash models appears to deal with the process of bargaining. Interestingly, the Nash outcome coincides with the outcome of a bargaining process proposed by Zeuthen (1930). On this, see Harsanyi (1956) who also shows the relationship between the Nash model and a suggestion of Hicks (1932).

Observation of labour–management bargaining indicates that agreements often are more costly than theory suggests. Strikes occur which impose costs on both sides even though both sides could have been better off by accepting the very same contract prior to a strike. Several directions are suggested in the literature in this regard. First, there are two-person cooperative game models in which offers and counter-offers are made until a settlement is reached. During this process, real time is assumed to elapse, and the total size of the

players' joint gain is supposed to shrink. Cross (1965) has such a model and, in an elegant paper, Rubinstein (1982) models the bargaining process in a way that turns the bargaining game into a noncooperative game having the Nash (1950) solution as its outcome. Second, it can be assumed, following Harsanyi and Selten (1972), that each player is ignorant of the payoff function of the other player. Each player makes a *demand*; however, if the two demands taken together lie beyond the payoff possibility frontier of the game, then no agreement is made. A third line of investigation, not formally applied to bilateral monopoly, is that of *repeated games* or *supergames*. Under this approach, one instance of bargaining between two players is seen as one episode in a larger game. For example, in labour–management negotiations, a contract is reached for a specific interval, say three years, and both firm and union are concerned with the effect that the current contract will have on later contract negotiations. This situation is easily seen as a game of many players, if it is added that the union may deal with more than one firm, and that the various contracts are inter-connected. This latter consideration embeds a bilateral monopoly in a larger context; hence may be thought to go beyond the present topic. Additional discussion of bargaining models can be found in Roth (1979) and Friedman (1986).

See Also

- ▶ Bargaining
- ▶ Cournot, Antoine Augustin (1801–1877)
- ▶ Game Theory
- ▶ Nash Equilibrium
- ▶ Zeuthen, Frederik Ludvig Bang (1888–1959)

Bibliography

Bowley, A. 1928. Bilateral monopoly. *Economic Journal* 38: 651–659.

Cross, J. 1965. A theory of the bargaining process. *American Economic Review* 54: 67–94.

Edgeworth, F. 1881. *Mathematical psychics*. London: Kegan Paul.

Fellner, W. 1949. *Competition among the few*. New York: Knopf.

Friedman, J. 1986. *Game theory with applications to economics*. New York: Oxford University Press.

Harsanyi, J. 1956. Approaches to the bargaining problem before and after the theory of games. *Econometrica* 24: 144–156.

Harsanyi, J., and R. Selten. 1972. A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science* 18: 80–106.

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

Marshall, A. 1890. *Principles of economics*. 9th (variorum) edn, with annotations by C.W. Guillebaud. London: Macmillan, 1961.

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

Nash, J. 1953. Two person cooperative games. *Econometrica* 21: 128–140.

Roth, A. 1979. *Axiomatic models in bargaining*. Berlin: Springer.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

von Stackelberg, H. 1934. *Marktform und Gleichgewicht*. Vienna: Julius Springer.

Wicksell, K. 1925. Mathematical economics. In K. Wicksell, *Selected papers on economic theory*, ed. E. Lindahl. Cambridge, MA: Harvard University Press, 1958.

Zeuthen, F. 1930. *Problems of monopoly and economic warfare*. London: Routledge & Kegan Paul.

Bimetallism

Lawrence H. Officer

Abstract

A bimetallic monetary standard is a combination of two metallic standards, each of which could in principle stand alone. Bimetallism has advantages over monometallism; but can be an unstable system, with legal bimetallism becoming de facto monometallism. The Persian and Roman Empires practised bimetallism. England's de facto bimetallism was short-lived, and US bimetallism difficult to maintain. French bimetallism in 1815–73 stabilized the gold–silver market price ratio and also exchange rates among gold, silver, and bimetallic countries. Bimetallism ended in the 1870s.

Keywords
Bimetallic arbitrage; Bimetallism; Deflation; Gold standard; Gresham's law; Inflation; Latin Monetary Union; Market ratio; Mint ratio; Monetary base; Money supply; Monometallism; Seigniorage; Silver standard; Specie-flow mechanism

JEL Classifications
N2

A bimetallic monetary standard is a combination of two metallic standards, each of which could in principle stand alone, and often evolved into de facto monometallism.

The Nature of Bimetallism

Bimetallic metals are usually gold and silver, but there are exceptions. Ancient Rome was temporarily on a silver-bronze standard; in the 18th century Sweden and Russia experienced a silver–copper standard.

Under bimetallism, both gold and silver coins are full legal tender. The unit of account (dollar, franc, and so on) is defined in terms of a fixed weight both of pure gold and of pure silver. So there is a fixed legal (mint, coinage) gold–silver price ratio: number of grains or ounces of silver per grain or ounce of gold. Both gold and silver enjoy free coinage (the government prepared to coin bars of either metal deposited by any party) and are full-bodied (have legal or face-value equal to metallic value). Token subsidiary (always silver) coins can exist. Subsidiary coins are fractions of (have face value less than) the unit of account; token coins have face value less than metallic (inherent) value, and invariably have restricted legal-tender power. Token coins were not adopted by bimetallic countries until late in their experience with bimetallism, and in conjunction with the process of terminating that standard.

Private parties may melt, import, and export coins (domestic or foreign) of either metal. There is no restriction on non-monetary uses of the monetary metals. Paper currency and deposits may exist; they are convertible into legal-tender coins, either directly or via government-issued paper currency (itself directly convertible into coin). Both private parties and the government may choose the metallic coin, or mixture of coins, in which to discharge debt (including paper currency). However, a private party does not have the right to a direct governmental exchange of gold for silver, or silver for gold. Logically, though, domestic gold and silver coin would exchange privately at the mint ratio.

Advantages and Disadvantages of Bimetallism

Bimetallism has four advantages. First, it embodies two sets of coins – one from a metal with a high value–weight ratio (gold), the other from a metal with a low ratio (silver). These provide a medium of exchange for a wide range of economic transactions. The range can be extended in both directions: upper, via paper currency and deposits; lower, via token subsidiary coins. Neither is incompatible with a bimetallic standard. Second, as does a monometallic standard, the bimetallic standard provides a constraint on the money supply and therefore inflation; for the legal-tender coins constitute the monetary base (given government-issued legal-tender paper, perhaps the 'super monetary base'), and the government must acquire one or the other metal to increase the base. Because there is coinage on demand, there is also a check on reduction to the monetary base, and on deflation. Third, a bimetallic country or bloc of countries accommodates shocks so that resulting effects on monometallic countries' money supplies are dampened. This is done by stabilizing the gold–silver price ratio ('market ratio') on the world market, the bullion market, where non-monetary gold and silver (generally bars) are traded either among themselves or individually for some important currency. Fourth, in stabilizing the market gold–silver price ratio, the bimetallic country or bloc also stabilizes the exchange rates between 'gold currencies' and 'silver currencies'. Otherwise, these exchange rates would fluctuate,

defeating one of the usual purposes of metallic standards.

The alleged disadvantage of bimetallism (relative to monometallism) is that it is unstable. Suppose the bimetallic-country's mint ratio initially is in the neighbourhood of the market ratio. A shock in the world supply of one metal can change the market ratio so that the mint ratio is now outside its neighbourhood. If the resulting market ratio is above (below) the mint ratio, then silver (gold) is 'bad' money, overvalued at the mint; domestic payments will tend to be made in that, relatively cheaper, coin rather than gold (silver), the 'good' money, undervalued at the mint and relatively expensive in the market. Good money will tend to be exported to settle balance-of-payments surpluses, bad money imported to finance balance-of-payments deficits. If the divergence between the market and mint ratio is large, 'bimetallic arbitrage' occurs, whereby good money is melted and traded on the bullion market for the bad metal, and the bad metal imported to be coined. In both situations, Gresham's law is operative: bad money drives out good.

Given sustained payments imbalances and/or a large and persistent divergence between the market and mint ratio, bad-money monometallism results. (The good money may be eliminated from the money supply, or circulate at a market-determined value – available only at a premium.) To avoid this, the mint ratio could be altered to remain in conformity with the market ratio. If the mint ratio is under-corrected, monometallism is not stemmed; if the mint ratio is over-corrected, monometallism in the opposite metal can occur. Successive changes in the market ratio can lead to alternating effective gold monometallism and silver monometallism, under the rubric of legal bimetallism. There are costs to such an alternating monetary standard; there are also costs in periodically altering the mint ratio.

Theories of Bimetallic Stabilization

Stabilizing bimetallic arbitrage occurs as follows. Suppose a shock occurs, new gold discoveries, that decrease the market ratio: the market price of non-monetary gold falls relative to silver. The market ratio now is below the mint ratio, so gold is 'bad' (overvalued) and silver 'good' (undervalued) money. Silver leaves the monetary system to be sold in the world (bullion) market, with gold purchased with the proceeds and coined. First, the arbitrageurs make a profit: the value of the gold coins they obtain is greater than the value of the silver coins they initially sold. Second, there is increased supply of silver (the appreciated metal) and increased demand for gold (the depreciated metal) in the bullion market – the two transactions constituting one arbitrage transaction. The result is an increase in the market ratio, which rises toward the mint ratio. Thus, the incentive for the arbitrage is eliminated. Third, the composition of the money supply of the bimetallic country changed, with a higher proportion of gold to silver. The bimetallic country stabilized the market ratio (and incidentally the exchange rates between gold and silver currencies), via the endogenous gold–silver composition of its money supply.

This mechanism is effective only to the extent that the bimetallic country has sufficient stock of the undervalued metal to return the market ratio close to the mint ratio, so that the incentive to arbitrage vanishes before monometallism in the overvalued metal results. However, the situation is not so dire, because costs of arbitrage imply 'gold–silver price–ratio' points that define a band for the market ratio within which the ratio can fluctuate without triggering bimetallic arbitrage. If the bimetallic-country's commitment to its mint ratio is absolutely credible, then stabilizing speculation exists within the bimetallic-arbitrage band, such that the market ratio turns away from its nearest bound and towards the mint ratio. The situation is analogous to stabilizing speculation within gold-point spreads, under the international gold standard.

Two other forces making for bimetallic stability have been suggested by Marc Flandreau. The first is 'metal-specific arbitrage' between the bullion and monetary markets. If a metal depreciates on the bullion market by more than coinage and associated costs, then owners of bars in that metal

will coin them in lieu of supplying them to the bullion market. If a metal appreciates by more than melting and associated costs of bringing that coined metal to the market, then holders of coin of that metal will melt them and supply them to the market. The reduced supply of the depreciated metal and increased supply of the appreciated metal act to return the market ratio towards the mint ratio. Unlike bimetallic arbitrage, these are independent transactions. Therefore the costs of metal-specific arbitrage are below the costs of bimetallic arbitrage, and the former provide a 'metal-specific band' located within the 'bimetallic arbitrage band.' So metal-specific arbitrage is a stabilizing mechanism that becomes operative before bimetallic arbitrage.

The second force involves the bimetallic country (France) transacting with a gold-currency country (England) and a silver-currency country (Germany). There are franc–sterling gold points, and franc–mark silver points. Expressing exchange rates as percentage deviations from parity and specie points in percentage terms, the franc/sterling–franc/mark exchange-rate differential (via triangular arbitrage) proxies the mark/sterling exchange rate. Also, implicit mark–sterling parity (via franc bilateral parities) corresponds to the mint ratio. On the assumption of no bilateral specie-point violations, the mark–sterling exchange rate has as upper (lower) bound the sum (negative *sum*) of the franc–sterling export (import) point and the franc–mark import (export) point. Now, the mark–sterling exchange rate is itself a good representation of the gold–silver market price ratio, because the Bank of England (Bank of Hamburg) supports, within a narrow band, a fixed sterling (mark) price of gold (silver). For the market ratio above the mint ratio (parity), so that silver is overvalued, the upper bound correctly involves exporting gold (sterling) and importing silver (marks). The gold–silver market price ratio has a bimetallic-arbitrage band that is approximately double the width of the franc–sterling and franc–mark bilateral specie-point spreads. Hence specie flows to settle and adjust payments imbalances occur prior to bimetallic arbitrage.

Suppose that a bimetallic country has lost all its undervalued ('good') metal, so it has become monometallic in its overvalued coinage. Nevertheless, Oppers (2000) shows that a bimetallic-arbitrage band could exist, given that there is a second bimetallic country with a different mint ratio. The two countries' mint ratios each constitute a bound to the market ratio, with, as usual, a market ratio beyond a bound giving rise to arbitrage that returns the market ratio to the band. For this mechanism to operate, both countries must actually or potentially have large amounts of both coined metals in their money stock, where 'large' means relative to shocks in the bullion market.

Bimetallism Prior to the 19th Century

The Persian Empire had the first bimetallic standard, with a mint ratio of $13\frac{1}{2}$ to 1 (all known mint ratios are in favour of gold) for a long time. This ratio undervalued silver relative to the ratio elsewhere, and presumably merchants took advantage of the price-ratio discrepancies in their regular dealings. The Roman Empire was often gold–silver bimetallic, but periodically debased the coinage. The likely reason was to increase seigniorage rather than to realign the mint ratio in conformity with the market ratio or the mint ratio in other lands. Until the mid-19th century, bimetallism was the legal standard in Europe (including England), though the mint ratio was often altered. Traditionally, the gold–silver price ratio was lower in China and India than in Europe.

England was legally on a bimetallic standard from the mid-13th century, when gold was first coined. The mint ratio was often changed. England was effectively on a silver standard until late in the 17th century, because the British mint ratio was generally below European gold–silver price ratios. Gold coins passed at a market price (in terms of the silver shilling) rather than face value, again indicative of a silver standard. In 1663 the (gold) guinea was coined, with a legal value of 20 (silver) shillings. The silver coins in circulation were in horrible condition, due in part to past debasement, in part to private clipping and sweating of the coins. So the market price of the guinea increased above 20 shillings – to as

much as 30 shillings – implying a gold–silver price ratio that effectively overvalued gold relative to Continental ratios. England was in process of switching from an effective silver to an effective gold standard.

In 1696 silver was recoined, so the coins became full-bodied again, and a ceiling (periodically reduced) was placed on the market price of the guinea. The result was that, for a brief period at the turn of the 18th century, England had effective bimetallism, with full-bodied coins of both metals in circulation. However, gold continued to be overvalued and silver undervalued; silver was exported, gold imported; and a de facto gold standard resulted. It became a de jure standard, via legislations restricting the legal-tender power of silver (1774) and effectively ending free coinage of silver (1816).

The Coinage Act of 1792 placed the United States on a legal bimetallic standard. The mint ratio (15 to 1) – selected because it was approximately the market ratio at the time – turned out to overvalue silver, because the market ratio increased. By 1823 gold had virtually gone from circulation, and an effective silver standard resulted. In 1834 Congress increased the ratio to 16.0022 (in 1837, revised slightly, to 15.9884). From 1834 to 1873, the world gold–silver price ratio was consistently below 16, so the new ratio overvalued gold, and an effective gold standard resulted. However, the export of full-bodied Mexican (silver) dollars and US subsidiary silver protected the circulation of underweight foreign silver pieces, which circulated at face value; so in a sense effective bimetallism continued. Only in the early 1850s, when the market gold–silver price ratio fell (due to gold discoveries and new production), did the United States begin to lose its remaining silver coins. In 1853, to retain the silver, Congress reduced subsidiary coins (below a dollar) to token status, with limited legal-tender power. The United States now was on a de facto gold standard. Legal bimetallism remained until 1873, when coinage of the silver dollar was terminated. One year later, silver was virtually demonetized; all silver coins (including the dollar) were restricted to maximum legal tender of five dollars in any payment.

Bimetallic France in the 19th Century

In 1803 France made the franc the monetary unit, and solidified and made effective the mint ratio of $15\frac{1}{2}$ that had been established in 1785. From the end of the Napoleonic Wars until 1873, while France retained that bimetallism, the market gold–silver price ratio remained in the neighbourhood of $15\frac{1}{2}$. (Also, exchange rates among gold, silver, and bimetallic countries were stable.) The stability of the market ratio was remarkable in the face of severe shocks to the bullion market. In the 1850s gold production increased tremendously due to gold discoveries in California and Australia, putting strong downward pressure on the market price ratio. In the 1860s gold production stopped increasing, and exploitation of Nevada silver discoveries put strong upward pressure on the ratio.

The steady market gold–silver price ratio was due primarily to the continued bimetallism of France, which acted as a buffer to shocks and thus stabilized the gold–silver market price ratio. What gave France this power were its large economic size, the substantial amounts of both gold and silver coins in its circulation, and its credible commitment to bimetallism at an unchanged mint ratio. Therefore, French bimetallic arbitrage operated – in the 1850s and early 1860s via gold imported and coined and silver melted and exported, in the later 1860s via the opposite activities. Stabilizing speculation within the bimetallic-arbitrage band, stabilizing bilateral specie flows, and metal-specific arbitrage were also elements in the French stabilization service. In 1865 the French stabilizing force was enhanced by formation of the Latin Monetary Union (LMU), in which France, Belgium, Switzerland, and Italy adopted a common bimetallism.

Some scholars, especially Oppers (1995, 2000), believe, rather, that France underwent serial monometallism, with bimetallism transformed to a de facto silver standard in the 1830s and 1840s, and the latter yielding to a de facto gold standard in the 1860s. Yet a parity band (with stabilizing speculation within the band) existed, with the French mint ratio the lower bound and the US mint ratio the upper bound in 1834–61, followed

subsequently by the French ratio the upper bound and the Russian ratio the lower bound. This interpretation of history is doubtful, for the strong propensity to use both metallic currencies was characteristic only of France. Also, Russia's mint ratio was inoperative at the time, as the country had an inconvertible paper currency.

In the early 1860s the future LMU countries, if not on a de facto gold standard, were certainly moving towards it. With the market ratio below the mint ratio, silver was being lost. To protect silver circulation, the individual countries made subsidiary coins token currency; while in 1866 the LMU came into effect, mandating reduction of the silver content and restriction of the legal-tender power of all silver coins except the largest, that is, the five-franc piece, which remained full-bodied.

French, LMU, and world bimetallism ended in the 1870s. The proximate cause was Germany's move to a gold standard, financed by the French indemnity that resulted from the Franco–Prussian War. Germany's release of silver put upward pressure on the gold–silver market price ratio. France was not prepared to accept the gold loss and silver inflow that would result from continued adherence to bimetallism. France (and Belgium) limited silver coinage in 1873, followed by the LMU mandating limits on coinage of the five-franc silver piece in 1874–6. In 1878 coinage of that piece was terminated. The existing five-franc coins retained full legal-tender power. France, along with Belgium and Switzerland, went on a 'limping' gold standard, redeeming government-issued paper money in either gold or silver at the discretion of the authority.

See Also

- ▶ Gold Standard
- ▶ Gresham's Law
- ▶ Silver Standard

Bibliography

Eichengreen, B., and M.R. Flandreau. 1996. Blocs, zones and bands: International monetary history in light of recent theoretical developments. *Scottish Journal of Political Economy* 43: 398–418.

Einzig, P. 1970. *The history of foreign exchange*. 2nd ed. London: Macmillan.

Flandreau, M.R. 1997. As good as gold? Bimetallism in equilibrium. In *Monetary standards and exchange rates*, ed. M.C. Marcuzzo, L.H. Officer, and A. Rosselli. London: Routledge.

Flandreau, M.R. 2002. 'Water seeks a level': Modeling bimetallic exchange rates and the bimetallic band. *Journal of Money, Credit and Banking* 34: 491–519.

Flandreau, M.R. 2004. *The glitter of gold*. Oxford: Oxford University Press.

Friedman, M. 1990. Bimetallism revisited. *Journal of Economic Perspectives* 4 (4): 85–104.

Gallarotti, G.M. 1994. The scramble for gold: Monetary regime transformation in the 1870s. In *Monetary regimes in transformation*, ed. M.D. Bordo and F. Capie. Cambridge: Cambridge University Press.

Leavens, D.H. 1939. *Silver money*. Bloomington/Indiana: Principia Press.

Martin, D.A. 1968. Bimetallism in the United States before 1850. *Journal of Political Economy* 76: 428–442.

Martin, D.A. 1973. 1853: The end of bimetallism in the United States. *Journal of Economic History* 33: 825–844.

Officer, L.H. 1996. *Between the dollar–sterling gold points: Exchange rates parity and market behavior*. Cambridge: Cambridge University Press.

Oppers, S.E. 1995. Recent developments in bimetallic theory. In *International monetary systems in historical perspective*, ed. J. Reis. Houndmills/London: Macmillan.

Oppers, S.E. 2000. A model of the bimetallic system. *Journal of Monetary Economics* 46: 517–533.

Redish, A. 1994. The Latin Monetary Union and the emergence of the international gold standard. In *Monetary regimes in transformation*, ed. M.D. Bordo and F. Capie. Cambridge: Cambridge University Press.

Redish, A. 2000. *Bimetallism: An economic and historical analysis*. Cambridge: Cambridge University Press.

Velde, F.R., and W.E. Weber. 2000. A model of bimetallism. *Journal of Political Economy* 108: 1210–1234.

Wilson, T. 2000. *Battles for the standard*. Aldershot: Ashgate.

Bioeconomics

Colin W. Clark

The word bioeconomics (sometimes bionomics) has been used to describe two separate fields of investigation: (a) the economics of biological systems – that is the ways in which biological

organisms and communities utilize scarce resources such as space, time, and sources of sustenance; and (b) biological resource economics – the ways in which the economic activities of human societies interact with the dynamics of biological systems. Bioeconomics is thus either a branch of biology (Wilson 1975; May 1981; Krebs and Davies 1984), or of economics (Clark 1976). For further discussion of the second interpretation. See ▶ Renewable Resources.

Two mathematical paradigms that play major, and complementary, roles in bioeconomics (both meanings) are optimization theory and the theory of competitive games. The philosophical basis for the use of optimization models in biology is the Darwinian theory of natural selection and evolution (Darwin 1859): biological organisms are hypothesized to evolve so as to maximize their 'fitness', meaning (roughly) their ultimate contribution to the gene pool. Explicit models, however, usually employ more primitive objective functions, such as expected rate of food intake, probability of survival, or number of progeny produced per breeding cycle. Optimization models have been applied to the study of many aspects of animal behaviour, including foraging strategy, territoriality, reproductive strategy and life histories.

Game-theoretic models are also implied by the Darwinian paradigm, given that the strategies employed by an individual organism will interact with those of its competitors, predators, parasites and mutualist organisms.

Early models of the 'struggle for survival' were based largely on ordinary differential equation systems, following pioneering work of Lotka (1925) and Volterra (1931). These macroecological models did not attempt to describe the strategical behaviour of individual organisms, however. Game-theoretic models were introduced formally into biology by J. Maynard Smith (e.g. 1982) and other theoretical biologists. Maynard Smith's concept of an evolutionary stable strategy (ESS) is related to the notion of Nash competitive equilibrium in the theory of games. In biological terms, a strategy is said to be an ESS if, once established in a population, it cannot be invaded by a rare alternative strategy. The ESS concept has proved useful in modelling many strategic situations, such as aggression, foraging and antipredation behaviour, mating systems, the evolution of sex, and so forth.

The phenomenon of altruistic behaviour was long considered a paradox for Darwinian theory. A related problem pertains to the natural regulation of animal population – how is it that populations do not regularly overrun their resource base in Malthusian fashion? Early theories of 'group selection' (Wynne-Edwards 1962) have been rejected for the most part – animals do not adopt altruistic strategies 'for the good of the species'. The more recent theory of kin selection (see Dawkins 1976), however, indicates that altruistic behaviour could evolve among closely related animals. An extreme case is found in insect societies, where all members of the group are essentially clones of the same queen. An alternative explanation of altruism is based on the concept of reciprocity in repeated games (Axelrod 1984).

The overpopulation problem is resolved by many species through the institution of territoriality. Explanations for non-territorial species remain somewhat unsatisfactory. Many highly fecund species do respond rapidly to ephemeral resource supplies, and then experience high mortality rates when the resource base declines. Among these so-called r-selected species occur many of the major agricultural pests.

See Also

- ▶ Lotka, Alfred James (1880–1949)
- ▶ Predator–Prey Models
- ▶ Volterra, Vito (1860–1940)

Bibliography

Axelrod, R.M. 1984. *The evolution of cooperation*. New York: Basic Books.

Clark, C.W. 1976. *Mathematical bioeconomics: The optimal management of renewable resources*. New York: Wiley–Interscience.

Darwin, C. 1859. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. London: John Murray.
Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford University Press.
Krebs, J.R., and N.B. Davies. 1984. *Behavioural ecology: An evolutionary approach*, 2nd ed. Oxford: Blackwell.
Lotka, A.J. 1925. *Elements of physical biology*. Baltimore: Williams and Wilkins.
May, R.M. (ed.). 1981. *Theoretical ecology: Principles and applications*. Oxford: Blackwell.
Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.
Volterra, V. 1931. *Leçons sur la théorie mathématique de la lutte pour la vie*. Paris: Gauthier-Villars.
Wilson, E.O. 1975. *Sociobiology: The modern synthesis*. Cambridge, MA: Harvard University Press.
Wynne-Edwards, V.C. 1962. *Animal dispersion in relation to social behaviour*. Edinburgh: Oliver and Boyd.

Biological Applications of Economics

Gordon Tullock

Both Darwin and Wallace, the two independent discoverers of biological evolution, specifically said that the idea came to them while reading Malthus's work on population. Since Malthus was history's first professor of Economics, this was clearly the most important influence of economics on biology. It is particularly interesting because Malthus's book on population has turned out to have relatively little predictive value in dealing with the human race in the roughly 150 years since it was written, but does fit non-human specie rather well. In a way he was a better biologist than an economist.

Surprisingly, after this promising start, to a large extent, economics and biology developed independently. Herbert Spencer made some use of evolution in his economic work, and other economists – Armen Alchian is the name that comes immediately to mind – have also made use of evolutionary ideas in economics. But until very recently there was almost no evidence of any biological concern with economics. There would be occasional articles in each of these disciplines which would show some minor contact with the other, but the phenomena was of the second or third order of smalls.

This comparative lack of cross stimulation was quite surprising granted the fact that both disciplines involve essentially the same intellectual construct, maximization subject to constraint. If one looks at present-day articles, in the *American Naturalist* and the *American Economic Review*, their superficial resemblance is quite high and their basic structure is also rather similar. In both cases, the standard article consists of application of optimizing methods to predict phenomena in the real world, and then statistical testing. Interestingly, in both cases, these articles normally perform their statistical tests on data which has been collected by other people. In both cases of course, a certain amount of direct data collection either by observation or experiment is present, but basically the dependence is on data provided by others.

In fact, the structural similarity between biology and economics is extremely strong. The evolutionary hypothesis in biology implies quite strongly that individual plants and animals 'act' as if they were attempting to maximize the frequency of their genes in the future. Of course, there is no genuine 'acting'. The dandelion for example, doesn't do anything much. Nevertheless, the selection process together with random changes in the genes, makes the dandelion more and more efficiently adapted to its environment. Of course the other species are also changing so that the environment is continuously changing. It is equivalent to firing at a randomly moving target.

Biologists regularly use language which might imply to the careless reader that animals and plants do consciously make plans and attempt to maximize. This is of course not what the biologist means. The process of actual selection itself functions as a mapping of what in human beings would be a set of conscious if not (as Micheal Ghiselin emphasizes) terribly intelligent decisions.

But although there has been some recent biological interest in economics, the present rather economic appearance of the biological journals is

I think an independent development. Once can find clear-cut examples of economic cross-stimulation. For a single example, a four-page note by the present author in the *American Naturalist* collected a grand total of 53 footnote citations in other parts of the biological literature (Tullock 1971). This was a simple economic explanation for the observed feeding habits of an English bird. Nevertheless, although there are other examples of the same kind of thing, most of the development was independent.

Evidence of this independence was the simple fact that although the general structure of articles in the two journals is very similar, there is an important stylistic difference. Economic articles usually take the form of a theoretical exposition which is entirely deterministic. Statistical theory is then brought in when it is tested with real world data. The biologist usually begins with probabilistic equations. The biological method is clearly more elegant, but also much harder. It is not obvious which is the most efficient research tool, but it is obvious that the biologists have not copied the economists in this area.

That there was a long period in which the two disciplines were operating with rather similar theoretical structures but with almost no cross-stimulation, requires an explanation. The most likely explanation seems to be that from let us say, 1860 until quite recently, most biologists were engaged in cataloguing and understanding the immensely diverse body of species in the world. No one knows exactly how many species there are; the number certainly exceeds ten million and biologists have devoted most of their attention to simply trying to find out what is out there. Darwin's book on barnacles (1851–4) was more typical of 19th- and early 20th-century biological research than his book on evolution.

The diversity of the biological world is almost unbelievable to the non-biologist. Even after a specie has been studied and described and entered into his reference books, the total number of such species is so immense that some may remain totally unknown even to experts in that field. E.O. Wilson, for example, is a very prominent biologist and would be so even if he had never written *Sociobiology* (1975). His special field is non-human societies. The mole-rat, a mammal with a life pattern rather similar to that of social insects, and the social spiders have been catalogued in the formal literature for over fifty years now. In *Sociobiology*, Wilson showed no signs of knowing they even existed. This is not in any sense a criticism of Wilson, but an indication of the real problem posed by the extraordinary diversity of the biological world. Had he decided to go through the entire literature and look for all social species, it would have taken many hundreds of lifetimes.

Be that as it may, there was relatively little contact between the two disciplines until recently, and although there is now more intellectual contact, it tends to be in certain rather applied fields, particularly environmental concerns. Garrett Hardin for example, a prominent biologist concerned with certain environmental problems, reinvented the economics of overgrazing, which he called 'The Tragedy of the Commons'; when it was called to his attention that this was essentially economic, he began a serious study of that aspect of economics. Since then he has worked with economists to produce joint projects in this general area (see, for example, Hardin and Baden 1977).

This is merely the most significant example of what is now quite a large body of cooperative research on such problems as pollution and environmental degradation. To a considerable extent the economic contribution to this joint research has amounted simply to pointing out that there are costs involved in preserving natural ecologies. Biologists tend to be extremely conservative in their approach to technology. The economist's role is frequently confined to pointing out that human welfare is also involved and suggesting trade-offs.

Another area where economists have for a long time been involved in biology is the specialized subdiscipline of agricultural economics. It should be said however, that in this case cross-fertilization has been rather minor. The basic objective of the professors of agricultural economics in our schools of agriculture has been to improve the returns of the farmers. For this purpose, they have engaged in applied economic

research in a number of fields. In general, however, they do not seem to have had any particular influence on the biological research which goes on in the same schools of agriculture. It should also perhaps be said here, that a great many of the agriculture economists devote their time to rationalizing economic subsidy programmes which although they certainly benefit farmers, injure everyone else.

The only other area of application is of course in the field of sociobiology. This has attracted a great deal of attention from economists and other social scientists and hence it is perhaps wise to emphasize here that it is currently only a minor field within the biological disciplines themselves. Nevertheless, it seems an obvious area for application of economics and such applications have been made.

The first problem here is that of territoriality which is frequently confused with the property relations which we find in human society. In fact, the biological specie have no guarantee of ownership and must one way or another defend their territory. The situation thus, is rather similar not to property ownership, but to competing retail establishments in a geographical area. The work of Losch (1937, 1938) is obviously relevant here and biologists have made good use of it. Indeed, the author of this note has contributed a couple of minor communications to the development of this area (Tullock 1979, 1983). The curious reader can find on page 272 of Wilson's *Sociobiology* a photograph of the Losch hexagons produced by a territorial species of fish.

As we move to more complex social structure it is more difficult to apply economics. Once again, Wilson used linear programming to study the distribution of Castes in the social insects. But an examination of the bibliography of his book will indicate that he was much more influenced by sociology than by economics in his general approach.

The dominance order, another important organizational structure found in the animal kingdom does not seem to have any direct analogies in economic reasoning. It is of course possible to apply economic analysis to the dominance order, but so far little progress been made along these lines. There is no reason to believe that economics has any comparative advantage here.

The complex societies of the social insects, the mole rats, possibly the social spiders, and certainly the sponges clearly are subject to economic analysis. All of them engage in complex cooperative activity which should be readily amenable to economic analysis. So far the opportunity has appealed to only one economist, the author of this item. Since his manuscript was never published, the field is open to any ambitious pioneer.

Micheal Ghiselin has undertaken a serious project to create an organization which will bridge the gap. So far he has been able to stimulate little interest in either discipline. This is not because of conscious opposition, but because most scholars find themselves too involved in their own discipline to take on the extra work. It is a particularly clear case of the narrow specialization which, unfortunately, dogs the learned professions.

Altogether, the amount of cooperation between economists and biologists is surprisingly small. In spite of similar roots and similar methods, the two disciplines have gone their own ways. In a few areas practical problems have brought them together, and there are occasional cases of the use of tools from one field in the other. This item covers economic applications in biology, but there are examples of reverse influence, the evolutionarily stable strategy, for example. Basically these influences are minor. I would like to say that the situation is changing and that there are signs of greater inter-disciplinary cooperation developing. Unfortunately, this would be to mislead the reader. I hope such developments will occur but the present signs are unfavourable. Economic analysis probably has a greater future in dealing with the communities of the social insects, but so far, little has been done in this area.

See Also

- ▶ Bioeconomics
- ▶ Competition and Selection
- ▶ Natural Selection and Evolution

References

Darwin, C. 1851 and 1854. *A monograph of the sub-class cirripedia, with figures of all the species*. 2 vols. London: Ray Society.

Darwin, C. 1851 and 1854. *A monograph of the fossil Lepadidae; or pedunculated Carripedes of Great Britain and a monograph of the fossil Balanidae and Verrucidae of Great Britain*. 2 vols. London: Palaeontographical Society.

Ghiselin, M.T. 1974. *The economy of nature and the evolution of sex*. Berkeley: University of California Press.

Hardin, G., and J. Baden. 1977. *Managing the commons*. San Francisco: W.H. Freeman and Company.

Losch, A. 1937. Population cycles as a cause of business cycles. *Quarterly Journal of Economics* 51: 649–662.

Losch, A. 1938. The nature of economic regions. *Southern Economic Journal* 5: 71–78.

Losch, A. 1964. *The economics of location* (trans: Woglom, W. H.). New Haven: Yale University Press.

Tullock, G. 1971. The coal-tit as a careful shopper. *The American Naturalist* 105: 77–80.

Tullock, G. 1979. On the adaptive significance of territoriality: Comment. *The American Naturalist* 113(5): 772–775.

Tullock, G. 1983. Territorial boundaries: An economic view. *The American Naturalist* 121(3): 440–442.

Wilson, E.O. 1975. *Sociobiology*. Cambridge, MA: Belknap.

Biology of Financial Market Instability

John Coates and Lionel Page

Abstract

Research in the biology of risk taking is today helping solve a problem identified in 1981 by Robert Shiller. In an influential article criticising the efficient markets hypothesis, Shiller demonstrated that 'measures of stock price volatility over the past century appear to be far too high – five to thirteen times too high – to be attributed to new information about future real dividends'. His paper has been debated ever since, but if it was pointing out a real phenomenon in 1981 then that point could be made even more forcefully today as the frequency and severity of market bubbles and crashes – in particular the housing bubble of 2002–07 and the credit crisis of 2008–09 – has only increased. How could biology help account for volatility of this magnitude and destructiveness?

Keywords

Biology; Bubbles; Crashes; Credit crisis; Crises; Efficient market hypothesis; Financial; Neuroeconomics; Preferences; Risk preferences

JEL Classifications

D84; D87; E3; G1

Many explanations have been proposed for the cycles of market bubble and crash – psychological herding, Minskean credit cycles etc. – but since the 2008–09 credit crisis a small number of researchers have turned their attention to a relatively under-researched phenomenon: time-varying risk aversion among the financial community. The suggestion here is that traders and investors could become more risk-seeking during a bull market, driving it into a bubble, and more risk-averse during a bear market, pushing it into a crash. In other words, risk preferences could shift pro-cyclically.

These researchers are thus taking issue with one of the most influential principles in economics, which states, in the words of George Stigler and Gary Becker, that 'one may usefully treat tastes as stable over time' (Stigler and Becker 1977). The principle of stable preferences was required in models of rational choice in order to ensure that choices were transitive; and it played an important role in limiting *ad hoc* explanations at a time when the psychological and physiological mechanisms underpinning economic preferences were not well understood. Today this assumption underlies many influential models of the financial markets. But it may be impairing our ability to understand market cycles; and recent data suggests it may be wrong.

For example, studies based on large databases drawn from brokerage accounts have found that during the housing bubble and credit crisis investors did indeed display pro-cyclical risk

preferences (Guiso et al. 2013; Smith and Whitelaw 2009; Malmendier and Nagel 2011). Models have also been developed in which risk preferences shift as a function of cycles in the financial markets (Campbell and Cochrane 1999; Verdelhan 2010). This growing literature on time-varying risk aversion in the financial community nonetheless leaves open many questions. What behavioural or neurological mechanisms could drive these variations in risk preferences? What is the magnitude of the changes? And crucially, if risk aversion varies, what conclusions can we draw regarding the market's ability to aggregate information efficiently?

In order to account for financial behaviour that has proved anomalous for existing theory, researchers have recently begun drawing on the protocols and findings of neuroscience, physiology and behavioural medicine. They have, for example, discovered many of the neural mechanisms involved in risk processing (Bossaerts 2009; Knutson and Bossaerts 2007; Kuhnen and Knutson 2005; Preuschoff et al. 2008); in attempts to second-guess competing investors (Bruguier et al. 2010); and in the distortions of judgement evident during bubbles (De Martino et al. 2013). A subset of this research has tried to identify the biological systems that shift risk preferences and thereby destabilise the markets. In what follows we review this subset of research on the biology of shifting risk preferences and financial market instability.

The Molecule of Irrational Exuberance

The research on the neurobiology of risk attitudes may be relatively new (Caplin and Schotter 2008), but it is based on substantial and decades-long research paradigms developed in physiology, neuroscience and behavioural medicine. These sciences have investigated, for example, how our physiology reacts to information and uncertainty (Hennessy and Levine 1979; Dickerson and Kemeny 2004; Pfaff 2006); how uncertain rewards can trigger a dopamine-mediated addiction to risk (Kuhnen and Knutson 2005; Berridge and Robinson 1998); how increases in anabolic hormones such as testosterone and growth hormone can increase a person's confidence and appetite for risk, even to pathological levels (Pope et al. 2000; van Honk et al. 2004; Reavis and Overman 2001); how chronic stress can alter our memory recall; and foster avoidance behaviour (Korte 2001; Sapolsky 2000; Kademian et al. 2005; Arnsten 2009).

The models developed in this research are today being extended into the financial realm in the hope of providing a scientific explanation for many risk-taking behaviours, ones that currently prove anomalous for existing economic theory (Coates et al. 2010). Important among these is the behaviour that during the internet bubble of the late 1990s came to be known as 'irrational exuberance'. Investors are said to be exuberant when they chase a bull market, buying more and more shares at ever more lofty evaluations, paying for price earnings multiples that cannot be justified by current earnings; and offering instead an unfounded optimism that the trend will continue indefinitely (Coates 2012). It is difficult to explain behaviour like this with the axioms of rational choice theory. But an explanation may be found in biology, in a remarkable phenomenon known as the 'winner effect'. It has been observed in both animals and humans that winning in a competition leads to increased risk-taking, which in turn can lead to further wins.

Biologists have found that an animal winning a competition or a fight for turf is statistically more likely to win the next agonistic encounter (Dugatkin and Druen 2004). The winner effect has been observed in a large number of species, from fish and reptiles to primates (Chase et al. 1994; Rutte et al. 2006). Studies of the winner effect have controlled for the competing animals' physical size (or what they term resource holding potential), motivation and aggression (Hurd 2006; Neat et al. 1998); but even with these controls in place, a pure winner effect emerges, suggesting that winning in itself contributes to future performance (Lehner et al. 2011). Once these empirical findings were established, biologists then began inquiring into the possible mechanism driving winner effects, and many were proposed: observable physical changes in a winning animal, such as increased pheromones,

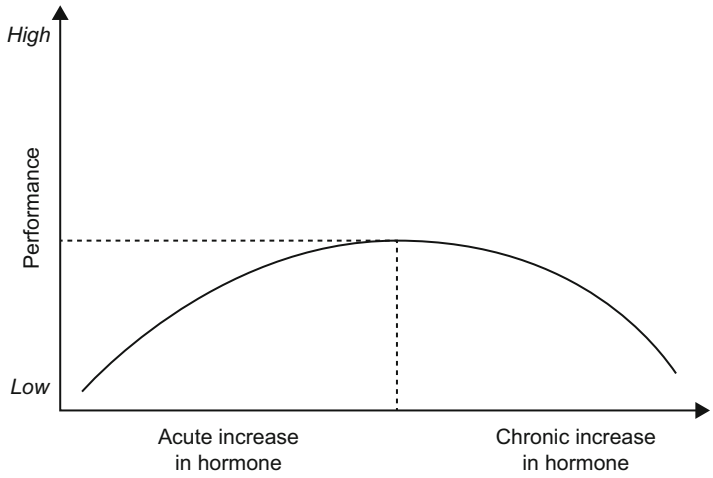
which would deter opponents from escalating new encounters (Rutte et al. 2006); winners revising up their estimates of their own abilities and deciding to escalate encounters (Dugatkin 1997; Mesterton-Gibbons 1999); or winners investing more effort in round-robin type competitions because they are closer to an overall victory (Konrad 2012; Konrad and Kovenock 2009; Malueg and Yates 2010).

However, the explanation that has received the most supporting data is one that focuses on the effects of competition on an animal's anabolic mechanisms (ones that build up tissues such as muscle), in particular on the naturally produced androgen hormone testosterone. Testosterone in an animal rises in anticipation of a competition (Wingfield et al. 1990) and rises still more after a victory (Trainor et al. 2004; Oyegbile and Marler 2005; Oliveira et al. 2009; Fuxjager et al. 2010, 2011), while falling after a defeat. Elevated levels of testosterone give an animal an edge in competition because it increases the animal's lean muscle mass, its haemoglobin and hence its blood's capacity to carry oxygen, as well as its confidence (Boissy and Bouissou 1994) and persistence (Andrew and Rogers 1972; Archer 1977). The winner effect may thus be driven by a physiological feedback loop in which winning leads to higher levels of testosterone, which in turn effectively increase the animal's resources, motivation, aggression and confidence, thereby raising the likelihood of further victories. This reaction may make sense from an evolutionary point of view: the loser of a fight is encouraged to retire into the bushes and nurse his wounds while the winner prepares for new challenges to his recently acquired rank.

Findings from animal studies can be extended to humans only with caution because the effects of physiological changes on our behaviour are mediated by a larger brain. Nonetheless, similar results haves been found in experiments with humans (Gladue et al. 1989). Athletes, for example, experience the same androgenic priming before a sporting contest and a further increase in testosterone after a win, a phenomenon observed for instance in tennis (Bateup et al. 2002) and wrestling (Elias 1981), as well as more purely cognitive contests such as chess (Mazur et al. 1992). Animals, including humans, harbour within themselves what amounts to a self-doping mechanism, giving them a shot of anabolic steroids when on a winning streak. Indeed it may be to trigger and harness the physiology of the winner effect that leads athletes, without knowing the biology involved, to 'psych themselves up' before matches by imagining victory or even by watching videos of previous victories (Carré and Putnam 2010).

Researchers have imported this biological model into the financial world by testing what they call 'the financial winner effect', hypothesising that physiological changes occur in traders and investors when they make above-average profits (Coates and Herbert 2008). They further extended the model by asking if at some point in the upward spiral of testosterone and victory the testosterone levels could become so elevated that they impair decision-making and risk-taking and lead to irrational exuberance. This extension of the winner effect model is based on a well-established phenomenon in pharmacology known as an inverted U-shaped dose–response curve (Fig. 1). What this means is that at very low levels of most hormones – adrenalin, cortisol, testosterone etc. – we perform badly at cognitive and physical tasks, but as the hormone increases so does our performance, leading to peak performance at the height of this curve. If, however, the hormone continues to rise then it can impair performance. By way of analogy, think, for example, of your morning cup of coffee: the first two cups may waken you and sharpen attention, but after five cups you may have difficulty focusing or even sitting still. You have gone over top of the dose–response curve.

In animals something like this has been observed with testosterone levels: that acute (i.e. moderate and short-lived) increases prove to be a powerful and effective means of empowerment, as in the winner effect; but if they continue to increase may morph effective risk-taking into dangerous behaviour. Animals with highly elevated testosterone tend to fight more, stray into the open more, neglect parenting duties, patrol areas that are too large and lose fat stores. As a result they suffer increased rates of predation and



Biology of Financial Market Instability, Fig. 1 Hormone dose–response curve

mortality (Beletsky et al. 1995; Dufty 1989; Marler and Moore 1988; Wingfield et al. 2001). At sufficiently high levels of testosterone, effective risk taking can morph into ill-considered and fatal risk-taking

Does something like this mechanism occur in traders and investors? Do risk takers in the financial markets experience a surge of testosterone when they make money, causing them to increase the size of their bets? Could this be the mechanism driving increasing levels of risk-taking during bull markets? Crucially, do the rising levels of testosterone cause traders to cross over the top of the dose–response curve and start placing bets in ever-increasing size with ever worsening risk–reward trade-offs until their bets go wrong and they lose more money than they made on the winning streak that originally fostered their 'irrational exuberance'?

Support for this hypothesis has come from a number of studies. In one study conducted on a trading floor in the City of London, hormones were sampled from 17 young male traders twice a day for a week and a half (Coates and Herbert 2008). It was found that these traders did indeed have significantly higher testosterone levels on days when they made an above-average profit. Were the profits causing the hormone change or the hormones causing the profits? The study design featured morning and afternoon sampling so it permitted the further observation that on days of high morning testosterone, the traders enjoyed an afternoon profit that was almost a full standard deviation higher than on 'low testosterone' days.

Other studies also looked at androgens and financial risk taking using a different marker of androgen exposure: the ratio of the second to fourth fingers (2D:4D) (Kondo et al. 1997; Malas et al. 2006; Manning et al. 1998). This marker is one of many physical traces left on our bodies by the levels of pre-natal androgen we were exposed to, much as a high-water mark is a trace of flood levels. One study looking at a cohort of 44 traders found that 2D:4D predicted their P&L as well as years of survival in the markets (Coates et al. 2009). Such a pattern in field data is backed up by experiments in the laboratory. In a financially motivated decision-making experiment, it was found that men and women with smaller digit ratios made riskier financial choices, and the effect was identical for men and women (Garbarino et al. 2011).

The findings of these studies present anomalous data for the efficient markets hypothesis. According to strong versions of this hypothesis, the market is random, so no trait, no skill, no training of a trader, not even their IQ, can improve their returns, any more than they could make a person better at tossing dice. But these findings present preliminary data suggesting that the levels of hormones can affect traders' P&L just as they affect performance among athletes. The question then becomes: how was the elevated testosterone

affecting P&L? Was it improving the traders' judgement? Their ability at predicting the market? Or was it increasing their risk appetite?

One study tried to answer this question (Coates and Page 2009). The researchers asked: are the androgen levels predicting the traders' skill as measured by their Sharpe Ratios, i.e., the ratio of their P&L to the variance of their P&L, or their risk? It was found that androgenic effects did not predict Sharpe Ratios but did predict risk, with higher levels of androgen exposure predicting higher levels of risk. Other more pharmacological studies have also found that increasing levels of testosterone increase appetite for risk (Apicella et al. 2014; Booth et al. 1999); and in still others that it encourages participants to choose the high-variance, low expected return decks of cards in the Iowa Gambling Task (Apicella et al. 2008; Pope et al. 2000; Reavis and Overman 2001; van Honk et al. 2004).

These findings – that testosterone does increase in traders when they experience an above-average P&L, and that testosterone increases risk appetite – suggest that a financial variant of the winner effect could be shifting risk preferences among the financial community during bull markets towards more risk seeking. Testosterone may be the molecule of irrational exuberance.

The Molecule of Irrational Pessimism

A different biological mechanism may contribute to the risk aversion that spreads during bear markets, frequently pushing them into a crash, a behaviour that has been termed 'irrational pessimism'. That mechanism is the stress response.

The stress response is often mistakenly taken to be a predominantly psychological phenomenon: the conscious feeling of being upset because something bad has happened or is expected to happen to you. But the stress response is more accurately understood as a physical preparation for impending movement. As such it includes changes in breathing, heart rate and blood pressure, and increasing levels of the stress hormones adrenalin and cortisol, both produced by the adrenal glands. The stress hormones suppress long-term functions of the body not needed during fight or flight, such as digestion and reproduction, and instead marshal fuel for immediate use: glucose from liver and muscles and, free fatty acids from fat cells. Adrenalin is a protein hormone with a short half-life in the blood (only a few minutes), while cortisol is a steroid hormone which, by triggering gene transcription, can exert long-term changes on almost all tissues of the body and brain.

The effects of cortisol differ dramatically between acute and chronic exposure, and they display the same inverted U-shape dose–response curve as testosterone. Acute stress is a normal part of life, and acute risks can even be enjoyable, as they are when playing sports or trading the markets. Acutely elevated cortisol in the brain interacts with dopamine, also called the pleasure, circuits. Rats will self-stimulate with cortisol. But chronic stress has very different effects, contributing to gastric ulcers, hypertension, immune disorders and blood glucose imbalances; while in the brain chronically elevated cortisol can promote anxiety, depression, learned helplessness, novelty avoidance and ambiguity aversion, and importantly it can affect memory recall, contributing to a selective attention to negative precedents (Erickson et al. 2003; Korte 2001). A chronically stressed person could therefore become more risk-averse.

Stress hormones, as part of our early warning system of potential threat, are highly sensitive to levels of novelty and uncertainty (Hennessy and Levine 1979). Novelty and uncertainty are endemic to financial markets. Indeed, the financial markets present a rare venue for researching stress because uncertainty can be measured objectively and accurately using the volatility of the markets. The VIX – an index of implied volatilities on US equities – is often called the Fear Index because it tracks uncertainty and stress in the financial system. In one study it was found that the stress response of traders was sensitively calibrated to levels of uncertainty and volatility in the market (Coates et al. 2008). As both historic and implied volatilities rose in the markets the traders were trading, so too did the levels of the traders'

cortisol. In this trading floor study, it was found that cortisol levels among traders rose 68% over a two-week period as volatility – and hence uncertainty increased.

This field work raised a crucial question: does the chronic elevation in stress affect the traders' risk aversion? In a follow-on study conducted in a research hospital, the authors used more a controlled experimental protocol to answer this question (Kandasamy et al. 2014). Using a placebo-controlled double blind crossover protocol, the authors raised pharmacologically the cortisol levels in volunteers a similar 68% over an eight-day period, to replicate the cortisol levels observed in the trader study; and through a computerised risk-taking task (implementing the Hey and Orme (1994) protocol to study risk preferences) it measured the utility and probability weighting functions underlying the participants' risk preferences. It was found that in response to the chronic increase in cortisol the participants risk aversion increased 44%. This was a large effect; and the conclusion suggested by the study is that risk preferences in the financial community are highly sensitive to sustained increases in volatility. Similarly, Cohn et al. (2015) show that simply priming traders with a situation of financial crisis leads them to be more risk-averse. The authors interpret such changes as being driven by physiology. Indeed it may have been this very physiological mechanism that contributed to the 'irrational pessimism' that afflicted the markets during the credit crisis of 2008–09, a period when implied volatilities rose to historically high levels.

Research into the physiological influences on financial market instability is in its infancy, but the research surveyed here suggests a new picture of financial risk-taking which departs from the assumption of stable risk preferences (Pearson and Schipper 2013). Risk-taking behaviour changes with alterations in our physiology; and our physiology is designed to sensitively calibrate our risk-taking to the amount of opportunity, uncertainty and threat in our environment. If the apparent opportunities of a bull market cause our endocrine systems to encourage more risk-seeking then, bull markets may segue into bubbles; and if the heightened uncertainty and losses of a bear market trigger a chronic stress response and promote risk aversion, then a bear market may spiral into a crash (Coates 2012; Coates and Herbert 2008).

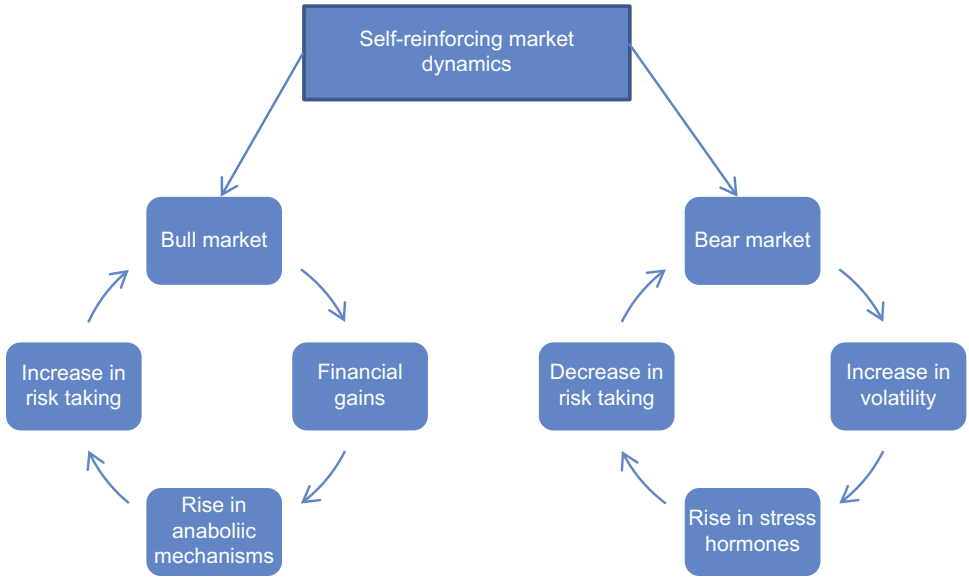
B

Conclusion

In his address to the 2015 American Economic Association meeting, Olivier Blanchard pointed out that one of the key insights from the credit crisis was the existence of 'dark corners' where feedback loops lead traditional linear macro models to fail. He encouraged researchers to investigate macro-finance models which accommodate such feedback loops. A key feature of the research on the biology of risk-taking is that it indeed supports self-reinforcing market dynamics (Fig. 2).

Recognising that biologically mediated shifts in risk preference can destabilise markets permits us to suggest novel policies for stabilising them. Market stability is served by having a diversity of opinions; and it may be served as well by having biological diversity among the traders and investors managing the world's wealth. How is this achieved? Androgens such as testosterone are higher in men than women (about five to ten times higher); and testosterone follows a pattern over the course of a man's life, rising to a peak in his 20s and falling thereafter, quite rapidly after the age of 50. If bull markets segue into bubbles partly due to rising androgen levels among the financial community, then perhaps bubbles are a young male phenomenon. And if so, then perhaps bull markets could be tamed if we had more women and older men managing money, because they may be less susceptible to the winner effect. Markets during crises could similarly benefit from having more women because their stress response differs from men. Women have stress hormones as high and as volatile as men, but research has found that their cortisol levels are less reactive to stressors stemming from a competitive situation (Stroud et al. 2002). They may thus be less susceptible to the spikes in risk aversion that help drive a bear market into a crash.

Keynes long ago invoked the notion of animal spirits to explain what we now call irrational exuberance and pessimism. Akerlof and Shiller



Biology of Financial Market Instability, Fig. 2 Self reinforcing market dynamics

(2010) have pointed out that the study of these animal spirits has been wrongly expelled from the study of economic phenomena. With the tools of biology and neuroscience, economists are now able to open the black box of these animal spirits, taking a first step towards taming them.

See Also

- Bubbles
- Bubbles in History
- Efficient Markets Hypothesis
- Experimental Methods in Economics
- Markets
- Neuroeconomics
- Rationality
- Speculative Bubbles
- Stock Price Volatility
- Uncertainty

Bibliography

Akerlof, G.A., and R.J. Shiller. 2010. *Animal spirits: How human psycholsogy drives the economy, and why it matters for global capitalism*. Princeton: Princeton University Press.

Andrew, R., and L. Rogers. 1972. Testosterone, search behaviour and persistence. *Nature* 237: 343–346.

Apicella, C.L., A. Dreber, B. Campbell, P.B. Gray, M. Hoffman, and A. Little. 2008. Testosterone and financial risk preferences. *Evolution and Human Behavior* 29(6): 384–390.

Apicella, C.L., A. Dreber, and J. Mollerstrom. 2014. Salivary testosterone change following monetary wins and losses predicts future financial risk-taking. *Psychoneuroendocrinology* 39: 58–64.

Archer, J. 1977. Testosterone and persistence in mice. *Animal Behaviour* 25(2): 479–488.

Arnsten, A.F.T. 2009. Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience* 10(6): 410–422.

Bateup, H.S., A. Booth, E.A. Shirtcliff, and D.A. Granger. 2002. Testosterone, cortisol, and women's competition. *Evolution and Human Behavior* 23(3): 181–192.

Beletsky, L.D., D.F. Gori, S. Freeman, and J.C. Wingfield. 1995. Testosterone and polygyny in birds. *Current Ornithology* 12: 1–41.

Berridge, K.C., and T.E. Robinson. 1998. What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 28(3): 309–369.

Boissy, A., and M. Bouissou. 1994. Effects of androgen treatment on behavioral and physiological responses of heifers to fear-eliciting situations. *Hormones and Behavior* 28(1): 66–83.

Booth, A., D.R. Johnson, and D.A. Granger. 1999. Testosterone and men's health. *Journal of Behavioral Medicine* 22(1): 1–19.

Bossaerts, P. 2009. What decision neuroscience teaches us about financial decision making. *Annual Review of Finance and Economics* 1(1): 383–404.

Bruguier, A.J., S.R. Quartz, and P. Bossaerts. 2010. Exploring the nature of 'trader intuition'. *Journal of Finance* 65(5): 1703–1723.

Campbell, J.Y., and J.H. Cochrane. 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107(2): 205–251.

Caplin, A., and A. Schotter. 2008. *The foundations of positive and normative economics: A handbook*. Oxford: Oxford University Press.

Carré, J.M., and S.K. Putnam. 2010. Watching a previous victory produces an increase in testosterone among elite hockey players. *Psychoneuroendocrinology* 35(3): 475–479.

Chase, I.D., C. Bartolomeo, and L.A. Dugatkin. 1994. Aggressive interactions and inter-contest interval: How long do winners keep winning? *Animal Behaviour* 48(2): 393–400.

Coates, J.M. 2012. *The hour between dog and wolf: How risk-taking transforms us, body and mind*. New York: Penguin-Random House.

Coates, J.M., and J. Herbert. 2008. Endogenous steroids and financial risk taking on a London trading floor. *Proceedings of the National Academy of Sciences* 105(16): 6167–6172.

Coates, J.M., and L. Page. 2009. A note on trader Sharpe Ratios. *PloS one* 4(11): e8036.

Coates, J.M., M. Gurnell, and A. Rustichini. 2009. Second-to-fourth digit ratio predicts success among high-frequency financial traders. *Proceedings of the National Academy of Sciences* 106(2): 623–628.

Coates, J.M., M. Gurnell, and Z. Sarnyai. 2010. From molecule to market: Steroid hormones and financial risk-taking. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365(1538): 331–343.

Cohn, A., J. Engelmann, E. Fehr, and M.A. Maréchal. 2015. Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review* 105(2): 860–885.

De Martino, B., J.P. O'Doherty, D. Ray, P. Bossaerts, and C. Camerer. 2013. In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron* 79(6): 1222–1231.

Dickerson, S.S., and M.E. Kemeny. 2004. Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin* 130(3): 355–391.

Dufty, A.M. 1989. Testosterone and survival: A cost of aggressiveness? *Hormones and Behavior* 23(2): 185–193.

Dugatkin, L.A. 1997. Winner and loser effects and the structure of dominance hierarchies. *Behavioral Ecology* 8(6): 583–587.

Dugatkin, L.A., and M. Druen. 2004. The social implications of winner and loser effects. *Proceedings of the Royal Society of London. Series B: Biological Sciences 271*(Suppl 6): S488–S489.

Elias, M. 1981. Serum cortisol, testosterone, and testosterone-binding globulin responses to competitive fighting in human males. *Aggressive Behavior* 7(3): 215–224.

Erickson, K., W. Drevets, and J. Schulkin. 2003. Glucocorticoid regulation of diverse cognitive functions in normal and pathological emotional states. *Neuroscience & Biobehavioral Reviews* 27(3): 233–246.

Fuxjager, M.J., R.M. Forbes-Lorman, D.J. Coss, C.J. Auger, A.P. Auger, and C.A. Marler. 2010. Winning territorial disputes selectively enhances androgen sensitivity in neural pathways related to motivation and social aggression. *Proceedings of the National Academy of Sciences* 107(27): 12393–12398.

Fuxjager, M.J., T.O. Oyegbile, and C.A. Marler. 2011. Independent and additive contributions of postvictory testosterone and social experience to the development of the winner effect. *Endocrinology* 152(9): 3422–3429.

Garbarino, E., R. Slonim, and J. Sydnor. 2011. Digit ratios (2D:4D) as predictors of risky decision making for both sexes. *Journal of Risk and Uncertainty* 42(1): 1–26.

Gladue, B.A., M. Boechler, and K.D. McCaul. 1989. Hormonal response to competition in human males. *Aggressive Behavior* 15(6): 409–422.

Guiso, L., P. Sapienza, and L. Zingales. 2013. *Time varying risk aversion*. Cambridge MA: National Bureau of Economic Research.

Hennessy, J.W., and S. Levine. 1979. Stress, arousal, and the pituitary-adrenal system: A psychoendocrine hypothesis. *Progress in Psychobiology and Physiological Psychology* 8: 133–178.

Hey, J.D., and C. Orme. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6): 1291–1326.

Hurd, P.L. 2006. Resource holding potential, subjective resource value, and game theoretical models of aggressiveness signalling. *Journal of Theoretical Biology* 241(3): 639–648.

Kademian, S.M., A.E. Bignante, P. Lardone, B.S. McEwen, and M. Volosin. 2005. Biphasic effects of adrenal steroids on learned helplessness behavior induced by inescapable shock. *Neuropsychopharmacology* 30(1): 58–66.

Kandasamy, N., B. Hardy, L. Page, M. Schaffner, J. Graggaber, A.S. Powlson, P.C. Fletcher, M. Gurnell, and J. Coates. 2014. Cortisol shifts financial risk preferences. *Proceedings of the National Academy of Sciences* 111(9): 3608–3613.

Knutson, B., and P. Bossaerts. 2007. Neural antecedents of financial decisions. *Journal of Neuroscience* 27(31): 8174–8177.

Kondo, T., J. Zákány, J.W. Innis, and D. Duboule. 1997. Of fingers, toes and penises. *Nature* 390(6655): 29.

Konrad, K.A. 2012. Dynamic contests and the discouragement effect. *Revue d'Économie Politique* 122(2): 233–256.

Konrad, K.A., and D. Kovenock. 2009. Multi-battle contests. *Games and Economic Behavior* 66(1): 256–274.

Korte, S.M. 2001. Corticosteroids in relation to fear, anxiety and psychopathology. *Neuroscience and Biobehavioral Reviews* 25(2): 117–142.

Kuhnen, C.M., and B. Knutson. 2005. The neural basis of financial risk taking. *Neuron* 47(5): 763–770.

Lehner, S.R., C. Rutte, and M. Taborsky. 2011. Rats benefit from winner and loser effects. *Ethology* 117(11): 949–960.

Malas, M.A., S. Dogan, E.H. Evcil, and K. Desdicioglu. 2006. Fetal development of the hand, digits and digit ratio (2D: 4D). *Early Human Development* 82(7): 469–475.

Malmendier, U., and S. Nagel. 2011. Depression babies: Do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics* 126(1): 373–416.

Malueg, D.A., and A.J. Yates. 2010. Testing contest theory: Evidence from best-of-three tennis matches. *Review of Economics and Statistics* 92(3): 689–692.

Manning, J.T., D. Scutt, J. Wilson, and D.I. Lewis-Jones. 1998. The ratio of 2nd to 4th digit length: A predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Human Reproduction* 13(11): 3000–3004.

Marler, C., and M. Moore. 1988. Evolutionary costs of aggression revealed by testosterone manipulations in free-living male lizards. *Behavioral Ecology and Sociobiology* 23(1): 21–26.

Mazur, A., A. Booth, and J.M. Dabbs Jr. 1992. Testosterone and chess competition. *Social Psychology Quarterly* 55: 70–77.

Mesterton-Gibbons, M. 1999. On the evolution of pure winner and loser effects: A game-theoretic model. *Bulletin of Mathematical Biology* 61(6): 1151–1186.

Neat, F.C., F.A. Huntingford, and M.M.C. Beveridge. 1998. Fighting and assessment in male cichlid fish: The effects of asymmetries in gonadal state and body size. *Animal Behaviour* 55(4): 883–891.

Oliveira, R.F., A. Silva, and A.V. Canário. 2009. Why do winners keep winning? Androgen mediation of winner but not loser effects in cichlid fish. *Proceedings of the Royal Society B: Biological Sciences* 276(1665): 2249–2256.

Oyegbile, T.O., and C.A. Marler. 2005. Winning fights elevates testosterone levels in California mice and enhances future ability to win fights. *Hormones and Behavior* 48(3): 259–267.

Pearson, M., and B.C. Schipper. 2013. Menstrual cycle and competitive bidding. *Games and Economic Behavior* 78: 1–20.

Pfaff, D.W. 2006. *Brain arousal and information theory.* Harvard: Harvard University Press.

Pope, H.G., E.M. Kouri, and J.I. Hudson. 2000. Effects of supraphysiologic doses of testosterone on mood and aggression in normal men: A randomized controlled trial. *Archives of General Psychiatry* 57(2): 133–140.

Preuschoff, K., S.R. Quartz, and P. Bossaerts. 2008. Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience* 28(11): 2745–2752.

Reavis, R., and W.H. Overman. 2001. Adult sex differences on a decision-making task previously shown to depend on the orbital prefrontal cortex. *Behavioral Neuroscience* 115(1): 196–206.

Rutte, C., M. Taborsky, and M.W. Brinkhof. 2006. What sets the odds of winning and losing? *Trends in Ecology & Evolution* 21(1): 16–21.

Sapolsky, R.M. 2000. Glucocorticoids and hippocampal atrophy in neuropsychiatric disorders. *Archives of General Psychiatry* 57(10): 925–935.

Smith, D.R., and R.F. Whitelaw. 2009. Time-varying risk aversion and the risk–return relation. *NYU Stern School of Business working paper*.

Stigler, G.J., and G.S. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67(2): 76–90.

Stroud, L.R., P. Salovey, and E.S. Epel. 2002. Sex differences in stress responses: Social rejection versus achievement stress. *Biological Psychiatry* 52(4): 318–327.

Trainor, B.C., I.M. Bird, and C.A. Marler. 2004. Opposing hormonal mechanisms of aggression revealed through short-lived testosterone manipulations and multiple winning experiences. *Hormones and Behavior* 45(2): 115–121.

van Honk, J., D.J.L.G. Schutter, E.J. Hermans, P. Putman, A. Tuiten, and H. Koppeschaar. 2004. Testosterone shifts the balance between sensitivity for punishment and reward in healthy young women. *Psychoneuroendocrinology* 29(7): 937–943.

Verdelhan, A. 2010. A habit-based explanation of the exchange rate risk premium. *Journal of Finance* 65(1): 123–146.

Wingfield, J., R. Hegner, A.M. Dufty, and G.F. Ball. 1990. The 'challenge hypothesis': Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *American Naturalist* 136: 829–846.

Wingfield, J.C., S. Lynn, and K.K. Soma. 2001. Avoiding the 'costs' of testosterone: Ecological bases of hormone-behavior interactions. *Brain, Behavior and Evolution* 57(5): 239–251.

Birmingham School

F. Y. Edegworth

Thomas Attwood is a signal example of good sense and general intelligence overborne by a futile monetary theory. He was the leader of the 'Birmingham School' who advocated high prices maintained by inflation of the currency. Attwood and his followers taught a lesson needed by some

of their contemporaries when they insisted on the hardship inflicted on debtors by a fall in general prices, or rise in the value of the monetary standard. But the extent of the evil was greatly exaggerated when the resumption of specie payments in 1819 was made responsible for almost every trouble which subsequently befell the kingdom – the agricultural distress in England, the turbulence of O'Connell in Ireland, or the 'Rebecca' riots in Wales. The argument directed against the resumption deserves particular attention. It was held that the depreciation of paper with respect to gold just before the resumption was much less than the appreciation of gold with respect to things in general which followed the resumption. 'That measure (which it was said would only effect a charge to the extent of 3 per cent) had imposed an additional burthen of of 25, 30, or 40 per cent) on every man in the community in all cases of deed, mortgage, settlement, or contact.' Prof. Walker appears inclined to ascribe some weight to this argument (*Money*, p. 388; *Money, Trade*, and *Industry*, p. 282).

Thomas Attwood's advocacy of monetary reform derived strength from his political influence. He was the founder of the 'Political Union' at Birmingham and took an active part in the agitation for parliamentary reform. It was believed that Attwood desired political reform principally as a means whereby to obtain a rectification of the currency. To that end he moved in the reformed parliament for a select committee to inquire into the causes of the general distress. This motion, like others which emanated from the Birmingham School, was lost.

Thomas Attwood was greatly assisted in this monetary crusade by his brother Mathias (born 1779, died 1851). Mathias's speech on the currency, 11th June 1822, is placed by Alison with Huskisson's speech on the other side, as 'containing all that ever has, or ever can be, said on the subject'. In 1830 Mathias proposed a double standard of silver and gold; at the rate of 152859/13640 lbs. of silver to 1 lb. of gold (Hansard, 1830, vol. xxv, pp. 102–145; Alison, *History of Europe*, 1815–52, vol. iv, ch. xxii. §32). Mathias, unlike his brother, was a Tory. He was a successful banker, co-founder of the National Provincial Bank of Ireland, the Imperial Continental Gas, and other companies. So little is business power alone a guarantee of sound economical theory.

There is an appreciative account of Thomas Attwood and the Birmingham School in a series of letters which were addressed to the *Midland Counties Herald* in 1843 by two Birmingham men (T.B. Wright and J. Harlow) signing themselves Gemini, and were republished in 1844 in the form of a book under that title. The title Gemini was appropriate according to Sir Robert Peel, for 'the efforts of no single writer are equal to the production of so much nonsense' (Speech on the Bank Charter, 1844). According to this par nobile 'the political economy of Mr. Attwood has this one great distinguishing feature, that it releases the nation from the thraldom of the heart-chilling doctrine of Malthus. The world is capable of multiplying its production to an almost unlimited extent; the governments of the world would have only to provide for the proper distribution of the productions, and the wants of all people will be supplied.' Such are the beneficent results of 'accommodating our coinage to man, and not man to our coinage' (Gemini, Letter 24). The cardinal tenets of the 'Birmingham economists' are compendiously stated at page 104, and again at page 285 of Gemini.

There is in the library of the British Museum a life of Thomas Attwood by his grandson, C.M. Wakefield, 'printed for private circulation'; which throws much light on the history of the Birmingham School. Mr. Wakefield does not profess to interpret his grandfather's views on currency.

J.S. Mill devotes a paragraph to the refutation of Attwood's theory of currency (*Political Economy*, book iii. ch. xiii. §4).

Birth-and-Death Processes

Yuji Ijiri

Birth-and-death processes offer a helpful tool in analysing the growth process and the resulting size distribution of entities. The size of an entity

is measured by the number of elements that belong to it. The birth-and-death process may comprise both the process by which elements are added to or deleted from the entity as well as to the process in which entities are added to or deleted from the population of entities. Examples of entities and their elements (stated in parentheses) are: cities (residents), firms (employees, customers, sales units, asset units), persons (income units), genera (species), authors (articles published), and words (appearances in a text).

Size distributions of entities have attracted attention because quite frequently empirical data show clear patterns that conform to the Pareto law, a linear relationship between the log of size of an entity and the log of rank of the entity, where rank is measured in such a way that the largest entity in the population has rank 1. Not only does the linearity hold well, especially for larger entities, but also the slope parameter changes little over time and over different regions from which data are taken. Birth-and-death processes help provide possible explanations for this observed regularity (Simon 1955; Steindl 1965; Singh and Whittington 1968).

Birth-and-death processes are, in their basic form, stochastic processes in which the system moves within an ordered set of states, E_0, E_1, E_2, ..., through a series of transitions from a state E_n to its adjacent state E_{n+1} or E_{n-1} ($n \geq 0$ and $n > 0$ respectively), under a given set of transition probabilities that depend only upon the system's current state (Markov processes). If the system is in state E_n at time t, the probability that, between t and $t + \Delta t$ (a) the transition $E_n \rightarrow E_{n+1}$ occurs is $\lambda_n \Delta t + o(\Delta t)$, (b) the transition $E_n \rightarrow E_{n-1}$ ($n > 0$) occurs is $\mu_n \Delta t + o(\Delta t)$, (c) more than one transition occur is $o(\Delta t)$, and (d) no transitions occur is $1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t)$.

The process is called a birth process if the probability of a transition $E_n \rightarrow E_{n-1}$ is zero for all $n > 0$. In particular, if λ_n in the birth process is a constant λ for all n, it becomes the Poisson process, which yields the Poisson distribution:

$$p_n(t) = e^{-\lambda t}(\lambda t)^n / n!, \qquad n = 0, 1, 2, \ldots, \quad (1)$$

where $p_n(t)$ is the probability that the system, starting from E_0 at $t = 0$, will be in state E_n at time t.

The above birth process assumes that the system has only a single birth mechanism operating independently of the state the system has attained. In many empirical systems such as those involving biological and economic populations, the expected number of births is often proportional to the size the system has attained. Hence an important special case of the birth process is when λ_n is set equal to λn, where λ is a constant. This leads to the negative binomial distribution:

$$p_n(t) = \binom{n-1}{a-1} e^{-a\lambda t}\left(1 - e^{-\lambda t}\right)^{n-a}, \quad n \geq a, \quad (2)$$

where a (>0) is the size of the population at $t = 0$. Similarly, the birth-and-death process in which $\lambda_n = \lambda n$ and $\mu_n = \mu n$ leads to the following distribution:

$$\begin{cases} p_n(t) = \sum_{j=0}^{\min(a,n)} \binom{a}{j} \binom{a+n-j-1}{a-1} \alpha^{a-j} \beta^{n-j} (1-\alpha-\beta)^j, & \text{for } n > 0, \\ p_0(t) = \alpha^a, \end{cases} \quad (3)$$

where

$$\alpha = \left\{\mu\left[e^{(\lambda-\mu)t} - 1\right]\right\} \Big/ \left[\lambda e^{(\lambda-\mu)t} - \mu\right]$$

and

$$\beta = \left\{\lambda\left[e^{(\lambda-\mu)t} - 1\right]\right\} \Big/ \left[\lambda e^{(\lambda-\mu)t} - \mu\right]$$

(Bailey 1964).

Each of the above systems involves only one entity whose distribution in size is at issue or multiple entities that all started at the same time and grow under the identical growth mechanism. But empirically interesting issues often relate to size distributions of entities that start at different times. Such analyses require not only a birth-and-death process of elements but also a birth-and-death process of entities to which elements belong.

If the birth process in which $\lambda_n = \lambda n$, leading to the probability distribution (Eq. 2), is applied to the birth process of elements and also to the birth process of entities which are all born with size 1 ($a = 1$), the resulting size distribution as $t \to \infty$, is given by:

$$p_n = \rho B(n, \rho + 1). \tag{4}$$

Here, $B(n, \rho + 1)$ is the beta function of n and $\rho + 1$;

$$B(n, \rho + 1) = \int_0^1 \tau^{n-1}(1 - \tau)^{\rho} \mathrm{d}t$$
$$= \Gamma(n)\Gamma(\rho + 1)/\Gamma(n + \rho + 1),$$
$$0 < n, 0 < \rho < \infty \tag{5}$$

where Γ is the gamma function and ρ is a parameter given by the ratio of the λ in the entity birth process to the λ in the element birth process. Since $\sum_{i=n}^{\infty} B(i, \rho + 1) = B(n, \rho)$, the distribution function, $F(n)$, cumulative from the right is:

$$F(n) = \sum_{j=n}^{\infty} f(n) = \rho B(n, \rho) = \rho\Gamma(n)\Gamma(\rho)/\Gamma(n + \rho). \tag{6}$$

Using a property of the gamma function, $\Gamma(n)/\Gamma(n + \rho) \to n^{-\rho}$ as $n \to \infty$. Thus for a large n, $f(n) \to \rho\Gamma(\rho + 1)n^{-(\rho+1)}$ and $F(n) \to \rho\Gamma(\rho)n^{-\rho}$, hence:

$$\log F(n) \sim \log\rho\Gamma(\rho) - \rho \log n, \quad \text{for large } n, \tag{7}$$

which shows a linear relation on the log-log scale between the size and rank ($F(n)$ times the number of entities in the population) of an entity.

The distribution given by Eq. 4 is called the Yule distribution (Yule 1924). Yule constructed the probability model to explain the distribution of biological genera by numbers of species. However, its wide range of applicability to empirical size distributions in various fields has been recognized, and its properties and variations have been analysed extensively.

At the heart of many 'contagious' phenomena is the so-called Gibrat law of proportionate effect, which says that the expected percentage growth rate in size is independent of the size already attained. The assumption, $\lambda_n = \lambda n$, in Eqs. 2, 3, and 4 incorporates this law. While Gibrat's law without new entries leads to the log-normal distribution, the law plus a constant rate of entry of unit-size entities has been shown to generate the Yule distribution (Simon 1955).

Simon and his colleagues have analysed many variations in the birth-and-death process that generate the Yule distribution and closely related distributions. In particular, the robustness of the Pareto law has been demonstrated under a variety of conditions. For example, (1) when all existing entities regardless of size have the same constant death rate, (2) when the entry rate of new entities is not a constant but a decreasing function of time, (3) when existing entities grow in proportion to their 'discounted size', the size whose components are discounted for the passage of time since birth so that entities with more recent growth have a better chance of growing than entities of comparable size but with an older growth history, and (4) when mergers and acquisitions are incorporated in the stochastic process (Ijiri and Simon 1977), the system's steady state is approximately Pareto.

The birth process under Gibrat's law of proportionate effect is also related to Bose-Einstein statistics, developed to describe the behaviour of physical particles. This statistics is in contrast to Maxwell-Boltzmann statistics under which all m^r arrangements of r elements in m entities have

equal probabilities of occurrence – an intuitively acceptable assumption. For example, each of the four arrangements of 2 elements, A and B, in 2 entities, (AB|), and (A|B), (B|A), and (|AB), has a 1/4 chance of occurrence. However, modern theory in statistical mechanics has shown that this statistics does not apply to *any* known particles. This observation led to the construction of Bose-Einstein statistics, under which all elements are *in* distinguishable and all distinguishable arrangements of r elements in m entities have equal probabilities of occurrence. Thus, in the above example, each of (**|), (*|*), and (|**) has a 1/3 chance of occurrence as against 1/4, 1/2, and 1/4, respectively, under Maxwell-Boltzmann statistics.

It can be shown that the birth process under Gibrat's law of proportionate effect preserves Bose-Einstein statistics at every iteration of growth. To illustrate briefly, consider two entities (|) each with size 1, where size is defined as *one plus* the number of stars it contains (alternatively the number of spaces delimited by a bar, a star, and/or a parenthesis). A star is thrown in at each iteration, each entity having a probability of receiving the star in proportion to its size. Hence, if the first star lands on the first entity (*|), its chance of receiving the next star becomes twice as large as the second entity. Thus, the probability of obtaining (**|) or (|**) is $1/2 \times 2/3 = 1/3$, leaving (*|*) the remaining 1/3 probability. These are the probabilities under Bose-Einstein statistics.

The birth process may include the birth of entities (throwing in bars) as well as the birth of elements (throwing in stars). For example, at each iteration a bar instead of a star may be chosen with a constant probability. If the bar is thrown in a space giving each space an equal chance of receiving it, an existing entity may be split by the bar, making it more difficult for an entity to attain a large size. The resulting size distribution is a geometric distribution. If the bar is placed only in a space adjacent to an existing bar, thus always creating an entity of a unit size, the resulting size distribution is the Yule distribution as discussed earlier.

Birth-and-death processes offer not only a tool to gain insight into the growth of economic entities but also a basis for policy evaluation on such matters as industrial concentration and mergers and acquisitions.

See Also

- ▶ Dynamic Programming and Markov Decision Processes
- ▶ Life Tables
- ▶ Pareto Distribution

Bibliography

Bailey, N.T.J. 1964. *The elements of stochastic processes*. New York: Wiley.

Ijiri, Y., and H.A. Simon. 1977. *Skew distributions and the sizes of business firms*. Amsterdam: North-Holland.

Simon, H.A. 1955. On a class of skew distribution functions. *Biometrika* 52(3/4): 425–440.

Singh, A., and G. Whittington. 1968. *Growth, profitability and valuation*. Cambridge: Cambridge University Press.

Steindl, J. 1965. *Random process and the growth of firms: A study of the Pareto law*. New York: Hafner.

Yule, G.U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions Series B* 213: 21–87.

Black, Duncan (1908–1991)

Bernard Grofman

Keywords

Black, D.; Borda, J.-C. de; Carroll, Lewis; Condorcet, Marquis de; Game theory; Majority rule; Median voter; Public choice; Single-peaked preferences; Spatial voting game; Voting paradoxes

JEL Classifications

B31

Born on 23 May 1908 in Motherwell, Scotland, Black studied at the University of Glasgow, where he received an MA (Mathematics and Physics) in 1929, an MA (Economics and Politics) in 1932, and a Ph.D. (Economics) in 1937. He also served there as Senior Lecturer in Social Economics, 1946–52. The bulk of his teaching career was at the University College of North Wales, Bangor: Lecturer in Economics, 1934–45; Professor of Economics, 1952–68; and Professor Emeritus 1968 onwards.

Black's very early research was in public finance, of which the major work is Black (1939). It is, however, his work in the 1940s and early 1950s (notably Black 1948a, b, c, 1949, 1950, and Black and Newing 1951), work which was integrated and expanded in Black (1958), which is the basis for his status as a father of the modern theory of public choice.

More than two centuries ago Condorcet (1785) demonstrated that majority rule need not yield a stable outcome when there are more than two alternatives to be considered. Although periodically rediscovered or reinvented by succeeding generations of scholars, the 'paradox of cyclical majorities' was, for all practical purposes, unknown to modern students of democratic theory until called to their attention by Duncan Black (see especially Black 1948a, 1958). Black demonstrated that the 'paradox' was not just a mathematical curiosity but rather was connected to important political issues such as manipulability of voting schemes (1958, p. 44; see also 1948a, p. 29) and the absence of strong similarity of citizen preference structures (Black 1958, pp. 10–14).

> Although Black was not the first to discover this phenomenon, his work is the foundation of all subsequent research on the problem. The investigations in this field of his principal predecessors, Condorcet and Lewis Carroll, had made no impact on the intellectual community of their day and had been completely forgotten. Their work is known today only because Black, after discovering the phenomenon himself, discovered his predecessors. (Campbell and Tullock 1965, p. 853)

Duncan Black's vision in the 1940s was a grand yet simple one: to develop a pure science of politics as a ramified theory of committees, so as to place political science on the same kind of theoretical footing as economics, with voters substituting for consumers. Because many of the basic ideas in his 1958 classic, *The Theory of Committees and Elections,* appear so 'obvious' in retrospect that it is hard to believe that they have not always been part of the stock of general human knowledge, and because this work understates by its silence the magnitude of Black's originality, the magnitude of Black's own contributions is often underappreciated. Black's great strength is that he has served as both synthesizer and pioneer. He rediscovered and reinterpreted for contemporary social science the strikingly modern probabilistic and game theoretic insights of long-dead theorists such as Dodgson (Lewis Carroll), Borda and Condorcet (for example, the paradox of cyclical majorities, the Condorcet criterion, the Borda criterion, optimizing strategies under the limited vote, results on manipulability of voting schemes, the Condorcet jury theorem); while himself developing such seminal ideas as single-peakedness, the importance of the median voter given ordinal preferences, and the notion of equilibrium in a spatial voting game (Black and Newing 1951; Black 1958, 1967, 1969, 1976). Black's work on Lewis Carroll (McLean et al. 1996) emphasizes Carroll's contributions to logic and the importance of his work on representation (under his real identity, that of the mathematician C.L. Dodgson) as a precursor to the modern theory of games and economic behaviour.

Underpinning virtually all of Black's work was the deceptively simple insight of modelling political phenomena in terms of the preferences of a given set of individuals in relation to a given set of motions, the same motions appearing on the preference schedule of each individual, where motions can be represented as points on a real line or in an N-dimensional space. Black's work on what (after him) has come to be called 'the theory of committees and elections' has been 'one of the pillars on which rests the contemporary theory of public choice' (Grofman 1981).

See Also

- ▶ Arrow's Theorem
- ▶ Borda, Jean-Charles de (1733–1799)

- ▶ Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794)
- ▶ Democratic Paradoxes
- ▶ Social Choice
- ▶ Social Choice (New Developments)
- ▶ Voting Paradoxes

Selected Works

1939. *The incidence of income taxes.* London: Macmillan; reprinted, New York: Kelley, 1965.

1948a. On the rationale of group decision making. *Journal of Political Economy* 56: 23–34.

1948b. The decisions of a committee using a special majority. *Econometrica* 16: 245–261.

1948c. The elasticity of committee decision with an altering size of majority. *Econometrica* 16: 262–270.

1949. The theory of elections in single-member constituencies. *Canadian Journal of Economics and Political Science* 14(2): 158–175.

1950, 1964. The unity of political and economic science. *Economic Journal* 60: 506–514. Repr. in *Game theory and related approaches to social behavior,* ed. M. Shubik. New York: Wiley, 1964.

1951. (With R.A. Newing.) *Committee decisions with complementary valuation.* Glasgow: William Hodge.

1958. *The theory of committees and elections.* Cambridge: Cambridge University Press.

1966. A simple theory of non-cooperative games with ordinal utilities. In *Papers on non-market decision making,* ed. G. Tullock. Charlottesville: Thomas Jefferson Center for Political Economy, University of Virginia.

1967. The central argument in Lewis Carroll's 'The Principles of Parliamentary Representation'. In *Papers on non-market decision making III,* ed. G. Tullock. Charlottesville: Thomas Jefferson Center for Political Economy, University of Virginia.

1969. Lewis Carroll and the theory of games. *American Economic Review* 59: 206–216.

1970. Lewis Carroll and the Cambridge Mathematical School of P.R.: Arthur Cohen and Edith Denman. *Public Choice* 8: 1–28.

1976. Partial justification of the Borda Count. *Public Choice* 28: 1–15.

Bibliography

Campbell, C.D., and G. Tullock. 1965. A measure of the importance of cyclical majorities. *Economic Journal* 75: 853–857.

Condorcet, N.C. de. 1785. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix*. Paris.

Grofman, B. 1981. The theory of committees and elections: the legacy of Duncan Black. In *Towards a science of politics: Essays in honor of Duncan Black*, ed. G. Tullock. Blacksburg: Public Choice Center, Virginia Polytechnic Institute and State University.

McLean, I., A. McMillan, and B.L. Munroe (eds.). 1996. *A mathematical approach to proportional representation: Duncan Black on Lewis Carroll*. Dordrecht: Kluwer.

Black, Fischer (1938–1995)

Perry G. Mehrling

Abstract

Fischer Black is best known for the Black–Scholes option pricing formula, which he regarded as an application of the capital asset pricing model (CAPM). He understood the CAPM as a model of general economic equilibrium and extended it from finance to macroeconomics, including the theory of money and the theory of business cycles. His work reveals that finance was the origin of the dramatic changes in macroeconomic thinking in the last quarter of the 20th century.

Keywords

American Finance Association; BDT (Black–Derman–Troy) term structure model; Black, F.; Black–Litterman model; Black–Scholes formula; Capital asset pricing model; Futures markets; Lucas, R.; Merton, R.; Noise trading; Option pricing theory; Prescott, E.; Real business cycles; Scholes, M.; Zero-beta model

JEL Classifications
B31

Fischer Black is best known for the eponymous Black–Scholes option pricing formula that laid the foundations for so much of modern finance (Black and Scholes 1973), a contribution that was recognized posthumously in the citation for the 1997 Nobel Prize in Economics that was awarded to Robert C. Merton and Myron Scholes. Today, the best known derivation of the famous formula follows the no-arbitrage argument laid out in Merton (1973), but Black approached the problem as simply an application of the capital asset pricing model (CAPM) developed by Sharpe (1964), Lintner (1965), and especially Jack Treynor (1962), whose version of CAPM was Black's first introduction to finance. Indeed, it is no exaggeration to say that not just the options formula but also everything Black ever wrote has its roots in CAPM, which Black always understood quite broadly as a model of general economic equilibrium, not just a model of how to price risky capital assets (Black 1972b).

Born 11 January 1938, Fischer Black grew up in Bronxville, New York, before attending both college and graduate school at Harvard University. After earning his Ph.D. in applied mathematics in 1964 for a thesis in the new area of artificial intelligence, Black took his first job as an analyst in the operations research section of the consulting firm Arthur D. Little, Inc. That's where he met Treynor and learned CAPM. Although he never took even a single course in either economics or finance, Black subsequently built a career as a financial consultant, a research professor (University of Chicago 1971–5, Massachusetts Institute of Technology 1975–83), and then a partner in the Wall Street investment firm Goldman Sachs (1984–95). He died prematurely on 30 August 1995, shortly after the publication of *Exploring General Equilibrium,* the book he considered to be his magnum opus.

Straddling the worlds of academia and business, Black developed his ideas by using practical problems in business as the stimulus for his abstract theorizing. The accessible early paper with Treynor, 'How to use security analysis to improve portfolio selection' (Treynor and Black 1973) set the agenda that would occupy Black and the generation of financial engineers that grew up after him, namely, to find practical applications of the new academic theories of finance. Just so, Black's early work with Myron Scholes for the Wells Fargo Bank sought to develop a new 'passive' portfolio strategy from the implications of CAPM, a kind of leveraged index fund that anticipated the later development of portfolio insurance (Black and Scholes 1974; Black 1988a; Black and Perold 1992). Similarly, his paper on 'Bank funds management in an efficient market' (1975) anticipated the eventual consequences of bank deregulation, and his paper 'Toward a fully automated stock exchange' (1971) anticipated the eventual consequences of computerized trading.

All of this was about remaking the world in the image of CAPM, an image that kept expanding in Black's mind as he worked to extend CAPM to a world without any riskless asset in his famous zero-beta model (1972a), to a world with long-term debt in the famous BDT term structure model (Black et al. 1990; Black 1995b), and to an international environment in his controversial universal hedging model (1974, 1990) that formed the analytical core of the Black–Litterman model of global asset allocation (Black and Litterman 1991, 1992).

The irony is that the world of the original CAPM is a world of debt and equity only, no options at all. That explains why Black was not sure that the opening in April 1973 of the Chicago Board Options Exchange was a good thing, even though it provided an immediate application for the Black–Scholes formula. Similarly, Black's extension of the options analysis to the problem of pricing commodity futures (1976), although immediately useful in the currency futures markets that sprang up after the collapse of the Bretton Woods fixed exchange rate system, left him unsure whether he was helping to move the world toward CAPM or away from it. From this point of view, his work on pension fund investment policy, the theory of business accounting, and a practical method of capital budgeting more clearly contributed to the creation of a CAPM world (1980b, a, 1993, 1988b).

Only after leaving academia for Goldman Sachs did Black come to fully appreciate the positive contribution of options and other derivatives to the brave new world of finance. The turning point was the theory of noise trading that he revealed for the first time in his presidential address to the American Finance Association (1986). Noise traders are people who trade, knowingly or not, without any information advantage. Earlier in his career, Black had assumed that such traders would eventually be driven out as markets become more and more efficient, but he changed his mind once he realized that 'Noise trading actually puts noise into prices'. As a consequence, 'we might define an efficient market as one in which price is within a factor of 2 of value; i.e. the price is more than half of value and less than twice value' (1986, 532–3). Because of noise trading, psychology matters for asset pricing, and it is in options prices that this effect can most clearly be seen; it shows up in the Black–Scholes formula as volatility.

Black's intellectual strategy to understand the world through the equilibrium lens of CAPM, as properly extended, was not confined to finance. He also used CAPM to lay the foundations of an alternative equilibrium understanding of macroeconomics, including the theory of money and the theory of business cycles, and he always considered this work at least as important as his work in finance. In this respect, his very first published paper, 'Banking and interest rates in a world without money: the effects of uncontrolled banking' (1970), set the agenda that would occupy him for the rest of his life. His two subsequent books *Business Cycles and Equilibrium* (1987) and *Exploring General Equilibrium* (1995a) had little impact on economics at the time they were published. In retrospect, however, they can be seen to have anticipated themes that eventually did enter economics, through the new classical revolution of Robert Lucas and his associates and the real business cycle revolution of Edward Prescott and his associates. More than anyone else, Fischer Black demonstrated that we must look to finance to discover the origin of the dramatic changes in macroeconomic thinking in the last quarter of the 20th century (Mehrling 2005).

See Also

- ▶ Accounting and Economics
- ▶ Capital Asset Pricing Model
- ▶ Efficient Markets Hypothesis
- ▶ Futures Markets, Hedging and Speculation
- ▶ Growth and Cycles
- ▶ Merton, Robert C. (Born 1944)
- ▶ Miller, Merton (1923–2000)
- ▶ Money and General Equilibrium
- ▶ Noise Traders
- ▶ Options
- ▶ Real Business Cycles
- ▶ Scholes, Myron (Born 1941)
- ▶ Term Structure of Interest Rates

Selected Works

1970. Banking and interest rates in a world without money: The effects of uncontrolled banking. *Journal of Bank Research* 1: 8–20. Reprinted as ch. 1 in Black (1987).

1971. Toward a fully automated stock exchange. *Financial Analysts Journal* 27, Part I: 28–35, 44; Part II: 24–28, 86–87.

1972a. Capital market equilibrium with restricted borrowing. *Journal of Business* 45: 444–445.

1972b. Equilibrium in the creation of investment goods under uncertainty. In *Studies in the theory of capital markets*, ed. M. Jensen. New York: Praeger.

1973 (With M. Scholes). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

1973 (With J. Treynor). How to use security analysis to improve portfolio selection. *Journal of Business* 46: 66–86.

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

1974 (With M. Scholes). From theory to a new financial product. *Journal of Finance* 19: 399–412.

1975. Bank funds management in an efficient market. *Journal of Financial Economics* 2: 323–339.

1976. The pricing of commodity contracts. *Journal of Financial Economics* 3: 167–179.

1980a. The magic in earnings: Economic earnings versus accounting earnings. *Financial Analysts Journal* 36: 19–24.
1980b. The tax consequences of long-run pension policy. *Financial Analysts Journal* 36: 21–28.
1986. Noise. *Journal of Finance* 41: 529–543.
1987. *Business cycles and equilibrium.* Cambridge, MA: Basil Blackwell.
1988a. Individual investment and consumption under uncertainty. In *Portfolio insurance: A guide to dynamic hedging*, ed. D. Luskin. New York: Wiley.
1988b. A simple discounting rule. *Financial Management* 17: 7–11.
1990. Equilibrium exchange rate hedging. *Journal of Finance* 45: 899–907.
1990 (With E. Derman and W. Toy). A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts Journal* 46: 33–39.
1991 (With R. Litterman). Asset allocation: Combining investor views with market equilibrium. *Journal of Fixed Income* 1: 7–18.
1992 (With R. Litterman). Global portfolio optimization. *Financial Analysts Journal* 48: 28–43.
1992 (With A. Perold). Theory of constant proportion portfolio insurance. *Journal of Economic Dynamics and Control* 16: 403–426.
1993. Choosing accounting rules. *Accounting Horizons* 7: 1–17.
1995a. *Exploring general equilibrium.* Cambridge, MA: MIT Press.
1995b. Interest rates as options. *Journal of Finance* 50: 1371–1376.

Bibliography

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.
Mehrling, P. 2005. *Fischer Black and the revolutionary idea of finance*. Hoboken: Wiley.
Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.
Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
Treynor, J. 1962. Toward a theory of market value of risky assets. In *Asset pricing and portfolio performance*, ed. R. Korajczyk. London: Risk Books.

Black–White Labour Market Inequality in the United States

Derek Neal

Abstract

During much of the 20th century, each successive generation of black Americans came closer to its white counterparts in terms of educational achievement and labour market success. This pattern of black–white progress has stalled since the mid-1980s. This chapter documents the current levels of black–white inequality in terms of human capital and labour market outcomes and then discusses factors that may sustain and perpetuate current levels of black–white inequality. It is much easier to understand the record of black–white progress during earlier decades than to understand the lack of progress in recent years.

Keywords

Affirmative action; Black–white educational achievement; Black–white labour market inequality in the United States; Black–white skill gap; Human capital investment; Incarceration; Inequality (explanations); Intergenerational transmission of human capital; Jim Crow South; Labour force participation rate; Labour market discrimination; Labour migration; Myrdal, G.; National Assessment of Educational Progress; National Center for Education Statistics; National Longitudinal Survey of Youth; Skill investment; Unemployment; Women's work and wages

JEL Classifications

D3

Gunnar Myrdal won a Nobel Prize in economics in large measure for path-breaking work that documented the magnitude and scope of black–white inequality in the United States prior to the Second World War. Blending social science with social commentary, Myrdal argued that the contrast between American ideals and the existing legal and social institutions that oppressed blacks created *An American Dilemma* (1944) that was moral and social as well as economic.

A Record of Progress

In subsequent decades, blacks have made much relative economic progress in the United States, but the pace of this progress has not been steady. For example, during the 1940s, 1960s, and 1970s the earnings of black men rose rapidly relative to those of white men, but this did not occur during the 1950s, 1980s, or 1990s. In fact, in recent decades the pace of relative economic progress for blacks has slowed and may be on the verge of stalling completely.

Table 1 presents data on the black–white earnings gap. The data come from the 1940–2000 decennial census files. Inconsistencies in the survey instrument as well as data-quality problems in some years make it difficult to create a consistent measure of hourly wages across census years. Here, I present data on annual labour earnings for workers who report working at least 48 weeks in the previous calendar year. For each year, numbers are given, separately for men and women, of the black–white ratio of average earnings and of the average percentile rank that black workers would have occupied in the white earnings distribution. I restrict the samples to ages 26–46 to minimize the number of lost of observations due to schooling or early retirement.

The results in Table 1 echo a common theme in the literature on black–white inequality. The 1960s and 1970s were decades when blacks made exceptional labour-market gains relative to whites both in terms of their position in the distribution of earnings and in terms of earnings levels. A significant literature debates whether government action during and after the civil rights era was a catalyst for black progress during the 1960s and into the 1970s. Smith and Welch (1989) and others emphasize the role of long-term improvements in the quantity and quality of black education as sources of black economic progress during the 20th century (see Card and Kruger 1992, 1996). While not disputing the importance of relative improvements in black education, Donohue and Heckman (1991) build a compelling case that federal government intervention did play a significant role in black progress during the civil rights era. They stress that black relative earnings rose significantly during the 1960s and 1970s within cohorts who were already adults at the beginning of these decades. They also note that black relative earnings rose during the 1960s primarily because of gains in the South, where civil rights laws were imposed on local communities by the federal government. Finally, they note that the decades- long wave of massive net black

Black–White Labour Market Inequality in the United States, Table 1 Black–white ratio of average annual earnings and average black percentile in the white earnings distribution

		Men		Women
Year	Ratio	Percentile	Ratio	Percentile
1940	0.45	0.167	0.39	0.126
1950	0.61	0.226	0.58	0.227
1960	0.60	0.214	0.63	0.268
1970	0.65	0.268	0.82	0.399
1980	0.73	0.343	0.97	0.494
1990	0.72	0.361	0.93	0.484
2000	0.70	0.367	0.88	0.464

Note: Data are from the Integrated Public Use Microdata Series (IPUMS) decennial census 1940–2000. The sample includes individuals between the ages of 26 and 45 who report positive wage and salary income and working at least 48 weeks in the previous calendar year. Sample weights 'slwt' are used for 1940 and 1950 and 'perwt' for 2000.

Black–White Labour Market Inequality in the United States, Table 2 (1) Fraction worked last calendar year (2) Fraction institutionalized

	White male age group				Black male age group			
Year of birth	26–30	31–35	36–40	41–45	26–30	31–35	36–40	41–45
1935–1939				0.938				0.820
				0.007				0.019
1940–1944			0.945				0.829	
			0.007				0.028	
1945–1949		0.947		0.927		0.822		0.779
		0.007		0.008		0.039		0.041
1950–1954	0.941		0.932		0.800		0.774	
	0.009		0.065		0.010		0.050	
1955–1959		0.933		0.888		0.756		0.709
		0.013		0.012		0.081		0.068
1960–1964	0.926		0.891		0.747		0.017	
	0.016		0.016		0.101		0.093	
1965–1969		0.898				0.715		
		0.018				0.116		
1970–1974	0.897				0.699			
	0.717				0.119			

Notes: Data for this table are from the decennial census IPUMS 1980–2000. The table displays the fraction of males who worked last year and fraction of males institutionalized. In order to be counted as working in the previous calendar year, a respondent must have (*a*) an affirmative, non-allocated response to the question 'Did this person work ...[during the previous calendar year]?' or (*b*) positive, non-allocated weeks worked or (*c*) positive non-allocated earned income or (*d*) positive, allocated weeks worked and a non-allocated indication of working since 1 January of the census year in question. Sample weights 'perwt' are used for 2000.

migration from the South to northern cities came almost to a complete stop around 1965. This one fact is strong prima facie evidence that the Civil Rights Act of 1964 did improve economic opportunity for blacks in the South.

Progress Stalled

The results in Table 1 also indicate that black economic progress since 1980 has been mixed at best. The male black–white earnings ratio fell slightly between 1980 and 2000, but black men did enjoy modest improvements in their relative position in the male earnings distribution over the 1980s and 1990s. (A dramatic increase in earnings dispersion over the period accounts for the different trends in these two measures of black–white earnings inequality among men.) Black women actually lost ground relative to white women according to both relative earnings measures over the 1980–2000 period.

However, it is not clear that black men fared better than black women relative to their white peers over this period. Neal (2004) points out that, even though black and white women have had similar labour force participation rates for several decades, racial differences in patterns of selection suggest that measured black–white earnings and wage gaps among women understate actual gaps in earnings opportunities. This bias arises because white women who do not work are more likely to be well-educated and married to a working spouse while black women who do not work are more likely to be single, less educated mothers receiving means- tested public assistance. The importance of this bias may have diminished since 1980 as government assistance to single mothers has decreased and the number of married career women has increased.

Further, the results in Table 1 are likely to overstate how well black men have fared relative to white men since 1980. Table 2 presents employment rates and institutionalization rates

for black and white men by age group and year of birth. Each diagonal row presents results from a particular census year, that is, 1980, 1990 or 2000. The employment rates refer to the past calendar year, and the institutionalization rates refer to the census date. Table 2 shows that the fraction of men who worked during the past calendar year has declined among both blacks and whites in recent decades (see Chandra 2000, for more details on patterns of male labour force participation by race). However, the rate of decline is much more dramatic among black men. By 2000, roughly 30 per cent of prime-age black men did not report any market work in the previous year. Further, in all age groups the relative decline in black employment rates is more than five percentage points. Thus, while Table 1 shows that black male workers continued to improve their position in the earnings distribution relative to working white men during the 1980–2000 period, it is not certain that black men continued to make relative gains in the distribution of potential earnings.

The most certain inference that one can draw from Table 2 is that the population of institutionalized black men has grown dramatically since 1980. In addition, since most institutionalized young adult men are incarcerated, Table 2 suggests that roughly one in ten black men aged 26–35 was housed in some type of prison or jail when the 2000 census was taken. (Neal 2006, shows that this rate is much higher among less-educated black men and dramatically lower among black college graduates.) Taken as a whole, Tables 1 and 2 suggest that black economic progress relative to whites has been anaemic at best since 1980.

Neal (2006) points out that, around 1990, black–white gaps in both educational attainment and achievement stopped closing among young adults and youth respectively. Thus, roughly since the mid-1980s, black youth and young adults have either barely kept pace or fallen farther behind their white peers with respect to numerous measures of human capital, such as achievement scores, total grade attainment, college graduation rates, and work experience. The National Assessment of Educational Progress, 2004, Long Term Trend scores provides some suggestive evidence that since 1999 black children have again begun to close the black–white gap in reading scores, but there is at best weak evidence of renewed progress in math. Overall, black–white math and reading gaps in 2004 among 9- and 13- year- olds are quite similar to the gaps observed in the late 1980s (NCES 2005).

The recent stability of black–white gaps in educational attainment and measured cognitive skills is an alarming development because the black–white skill gap is an important source of economic inequality between blacks and whites. Neal and Johnson (1996) and Johnson and Neal (1998) show that a large portion of black–white differences in earnings and wages can be accounted for by differences in basic reading and math skills among teenagers that pre-date labour market entry. Black–white skill gaps are a driving force behind black–white differences in labour market outcomes among adults for several reasons. First, the black–white skill gap among the current generation of adults is quite large. For example, respondents in the National Longitudinal Survey of Youth (NLSY), 1979, are in their forties now, and the black–white gap in Armed Forces Qualifying Test (AFQT) scores for this sample was over one standard deviation. (The black–white AFQT gap is smaller among youth tested as part of the NLSY, 1997, but the gap remains close to one standard deviation.) Second, measured labour market returns to skill are now at historical highs in the United States. Third, the current market gradients between labour market outcomes and various measures of human capital are even steeper for blacks than for whites. Black and white high-school dropouts, on average, experience markedly different labour market outcomes but, among persons with a college degree and strong reading and math skills, race is much less salient as a predictor of labour market outcomes (Neal 2006). Because the black–white skill gap is so costly to the current generation of black adults, economists are hard-pressed to explain the recent stability of the black–white skill gap. The 20th century saw several generations of black children make important human capital gains relative to their white peers during times when public expenditures on schooling and pre-school

programmes available to black communities were not nearly as high as they are now relative to comparable spending in white communities and when government did much less to ensure that skilled blacks would be treated fairly in the labour market as adults.

What Went Wrong?

This record of progress is a key starting place for discussing black–white inequality. The logic of basic models of the intergenerational transmission of human capital suggests that one should expect black–white skill convergence. Because the time and attention of each child is a fixed factor in the production of the child's human capital, there are decreasing returns to investments in any child. Thus, in the absence of spillover effects, any group of parents who are more skilled than some other group of parents by a factor k must invest more than k times as much in their children to maintain the same inter-group skill gap in the next generation. In many models, diminishing returns forces skill convergence between two groups unless there is a barrier that hinders investment among one group. The challenge for economists is to understand what barriers are present now in the black community that were not present during 1940–90.

Economists have put forth several theories concerning potential obstacles to skill investment by blacks. None fits all the facts. Coate and Loury (1993) described a model of statistical discrimination in which blacks do not invest because they expect employers to be less likely to reward them for investing. Employers do not see investment levels but rather a noisy signal of worker skill. Because employers believe that black workers are less likely to invest, they screen black workers more stringently, thus lowering the returns to black skill investments, as black workers anticipated. Further, the rational reluctance of black youth to invest confirms the beliefs of employers concerning black investment behaviour.

The Coate and Loury model has been quite influential because it provides an elegant theory of endogenous racial differences in human capital and labour earnings. However, the model is squarely at odds with a key feature of data on skills and labour market outcomes. As I note above, gradients between earnings and wages on the one hand and measures of achievement and attainment on the other are almost always as steep among blacks as among whites, and often steeper. This directly contradicts the scenario described in Coate and Loury (1993), and one cannot rescue their approach by arguing that the gradients observed in the data do not necessarily answer counterfactual questions concerning what less-skilled blacks would have earned if they had invested in skills. This model and others that explain statistical discrimination as a coordination failure are describing a market equilibrium and the resulting market gradients between skill and earnings in that equilibrium. However, no study has yet shown that there exists a gradient between any measure of labour market success and some dimension of worker skill that is systematically steeper among whites than among blacks in the post-civil rights era.

(Precise tests of the model are difficult because the skill in question should be observed by the econometrician but not by employers. Nonetheless, blacks do enjoy equal or greater measured returns to the measures of skill and attainment available in current data sets; see Neal 2006; Levy et al. 1995.)

A satisfactory explanation of the recent stagnation of black–white skill gaps must begin on the supply side by describing the factors that raise the cost of investing in skills within the black community. Recent work by Austen-Smith and Fryer (2005) provides a model of 'acting white'. In their model, loss of social cooperation constitutes an additional cost of human capital investment in the black community, and only the most gifted in the community actually invest. This model can produce the steep gradients that we observe between skills and both earnings and wages in the black community because blacks who invest in market skills enjoy expected returns from these investments that are high enough to offset any social sanctions they may suffer. However, the basic argument advanced by Austen-Smith and Fryer (2005) cannot account for all we know about

black–white skill differences. Their model is presented as a description of peer pressure, but black–white skill gaps are quite large when children begin school, widen during elementary school, and do not increase much if at all after students enter high school (Neal 2006). The gaps that exist prior to school entry are more likely to be connected to black–white differences in home environment than black–white differences in peer interactions. Further, it is not obvious why fears of being sanctioned for 'acting white' should have a more deleterious effect on black achievement during elementary school than during the teen years. Finally, if the social stigma of 'acting white' is sustaining the large black–white gaps in achievement and attainment that remain in 2005, we may need to think more carefully about potential sources of change in black culture during recent decades. It is logically possible but hard to imagine that the dramatic black progress observed during the 1940–90 period could have taken place in black communities where achievement and attainment were accompanied by sanctions for 'acting white'.

Because black–white skill gaps are quite large even among young children, it is natural to examine the roles of parents and families when trying to understand why recent cohorts of black children have failed to continue closing the black–white skill gap. Neal (2006) discusses changes in the wage structure and contemporaneous changes in family structure within the black community since 1980 that have reduced the resources available to children in black families. These changes may have adversely affected investment in black children, and if this is the case, the recent stability of the black–white skill gap will be temporary. Negative shocks to black wealth should only slow the process of black–white skill convergence. Even in models with imperfect credit markets, the standard expectation is that pure wealth effects will not persist indefinitely over generations. (See Loury 1981, and Mulligan 1997. Neal 2006, provides a detailed discussion of factors that influence black–white skill convergence.)

Recent studies of parenting behaviours do indicate that there are important black–white differences in ways that parents interact with children and that these differences contribute to black–white differences in cognitive development at an early age (see Brooks-Gunn, Duncan and Klebanov 1996; Brooks-Gunn et al. 1998), but it is not clear whether these parenting differences should be understood as differences in culture or differences in parenting practices that are driven by differences in family resources.

Conclusion

In closing, I must note that black workers may well face problems other than skill deficits. In particular, the extremely low earnings and employment levels currently observed among less-skilled black men may be more than the results of an interaction between low skill levels and economy-wide shifts in labour demand that favour skilled labour. Mailath, Samuelson and Shaked (2000) construct an informative model of discrimination against minority groups based on search behaviour, and in their model equilibria exist in which members of minority groups suffer wage discrimination and higher rates of unemployment because employers direct search effort to networks populated by majority group members. Because minority workers and firms know that employers are not searching in minority networks, minority workers have little bargaining power when they do create an encounter with an employer through their own search efforts. In this model, affirmative action policies that mandate colour-blind search eliminate inter-group wage differences because they give all workers the same bargaining power.

In light of the Mailath, Samuelson and Shaked model, consider the real possibility that skilled labour markets may be more heavily influenced by government anti-discrimination efforts. (There is suggestive evidence that this is the case; see Smith and Welch 1984; Leonard 1990. Further, Holzer, 1998, provides evidence that large firms, which tend to hire more skilled workers and use formal hiring methods, are significantly more likely to hire black workers than small firms.) If so, the forces identified by Mailath, Samuelson

and Shaked are a potential reason that less-skilled blacks fare so much worse relative to their white peers than highly skilled blacks. Further, the Mailath, Samuelson and Shaked reasoning helps us understand why gradients between skill and labour market outcomes have been relatively steep in the black community following the Civil Rights Act, but not before. (Welch 1973, was the first to note this reversal; see Neal 2006, for later results.)

Current black–white inequality is much less extreme than the inequality Myrdal observed, but the black–white inequality that remains is more ominous in some respects. The destitution of Southern blacks that Myrdal wrote about was clearly related to direct and oppressive action on the part of state and local governments that intentionally limited the educational and economic opportunities available to black citizens. Nonetheless, blacks made substantial economic and educational progress in the 1940s, and a combination of legal challenges and legislative efforts gradually began to undercut the systems of school financing and Jim Crow employment practices that afflicted blacks so greatly. In contrast, at the beginning of the 21st century blacks no longer face overt government oppression. Yet, since the mid- 1980s, black–white differences in potential wages and earnings have remained roughly constant or grown slightly, incarceration rates among black men have exploded, and black–white skill gaps have remained large and roughly constant. We still face *An American Dilemma,* but the primary causes of our current dilemma and the policy changes necessary to foster further progress are less clear than in Myrdal's day. The current experiences of blacks in the United States present a challenge for economists who wish to understand the dynamics of group outcomes within developed economies.

See Also

- ▶ Affirmative Action
- ▶ Inequality (Global)
- ▶ Wage Inequality, Changes in

Bibliography

Austen-Smith, D., and R. Fryer Jr. 2005. An economic analysis of acting white. *Quarterly Journal of Economics* 120: 551–583.

Brooks-Gunn, J., G. Duncan, and P. Klebanov. 1996. Ethnic differences in children's intelligence test scores: Role of economics deprivation, home environment and maternal characteristics. *Child Development* 67: 396–408.

Brooks-Gunn, J., M. Phillips, G. Duncan, P. Klebanov, and J. Crane. 1998. Family background, parenting practices, and the black–white test score gap. In *The Black–white Test Score Gap*, ed. C. Jencks and M. Philips. Washington, DC: Brookings Institution.

Card, D., and A. Kreuger. 1992. School quality and black–white relative earnings: A direct assessment. *Quarterly Journal of Economics* 107: 151–200.

Card, D., and A. Krueger. 1996. School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives* 10(4): 31–50.

Chandra, A. 2000. Labor-market dropouts and the racial wage gap: 1940–1990. *American Economic Review* 90: 333–338.

Coate, S., and G. Loury. 1993. Will affirmative-action policies eliminate negative stereotypes? *American Economic Review* 83: 1220–1240.

Donohue, J.I.I.I., and J. Heckman. 1991. Continuous versus episodic change: The impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29: 1603–1643.

Holzer, H. 1998. Why do small establishments hire fewer blacks than large ones? *Journal of Human Resources* 33: 896–914.

IPUMS (Integrated Public Use Microdata Series). Online. Available at http://www.ipums.org.usa/. Accessed 3 Feb 2006.

Johnson, W., and D. Neal. 1998. Basic skills and the black–white earnings gap. In *The black–white test score gap*, ed. C. Jencks and M. Philips. Washington, DC: Brookings Institution.

Leonard, J. 1990. The impact of affirmative action regulation and equal employment law on black employment. *Journal of Economic Perspectives* 4(4): 47–63.

Levy, F., R. Murnane, and J. Willett. 1995. The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics* 77: 251–266.

Loury, G. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49: 843–867.

Mailath, G., L. Samuelson, and A. Shaked. 2000. Endogenous inequality in integrated labor markets with two–sided search. *American Economic Review* 90: 46–72.

Mulligan, C. 1997. *Parental priorities and economics inequality.* Chicago: University of Chicago Press.

Myrdal, G. 1944. *An American dilemma*. New York: Harper and Brothers.

NCES (National Center for Education Statistics). 2005. Long-term trend major results. Online. Available at http://nces.ed.gov/nationsreportcard/ltt/results2004/. Accessed 19 Feb 2006.
Neal, D. 2004. The measured black–white wage gap among women is too small. *Journal of Political Economy* 112: S1–S28.
Neal, D. 2006. Why has black–white skill convergence stopped? In *The handbook of economics of education*, ed. E. Hanushek and F. Welch, Vol. 1. Amsterdam: North-Holland.
Neal, D., and W. Johnson. 1996. The role of pre–market factors in black–white wage differences. *Journal of Political Economy* 104: 869–895.
Smith, J.P., and F. Welch. 1984. Affirmative action in labor markets. *Journal of Labor Economics* 2: 269–301.
Smith, J., and F. Welch. 1989. Black economic progress after Myrdal. *Journal of Economic Literature* 27: 519–564.
Welch, F. 1973. Black–white differences in returns to schooling. *American Economic Review* 63: 893–907.

Blake, William (c1774–1852)

G. de Vivo

Blake was a member of the Political Economy Club from 1831 until his death, and in 1815–16 President of the London Geological Society (of which his friend Ricardo was also a member). His reputation as an economist was established by his 1810 tract, *Observations on the Principles Which Regulate the Course of Exchange; and on the Depreciated State of the Currency*, which came to be regarded as a standard work on the subject of foreign exchanges. He made the point that the actual (or *computed*) rate of exchange is determined by two different groups of causes. One, the demand and supply of foreign bills in the market, depending on the foreign payments the country has to make. The other, causes only affecting the value of the currency – i.e. the quantity and quality of metal in the coin, and the amount of currency compared with the commodities which have to be circulated by it. The rate of exchange as affected by the former causes he called the *real* exchange, the latter causes would instead affect what he called the *nominal* exchange. The combined effects of the two would determine the actual exchange (1810, p. 481). The distinction (which was then generally accepted: see e.g. Ricardo, *Works*, IV, p. 353) was not entirely well founded, however, because for instance changes in the price level caused by changes in the amount of money in circulation (Blake accepted the quantity theory) would affect the exports and imports of the country, and therefore could be seen as affecting the *real* exchange.

So far as currency questions are concerned, in this first work Blake adopted a straightforward bullionist position. He warned, however, on the dangers of deflation, and insisted that great caution be taken during the return to cash payments (1810, pp. 549 ff.).

In 1823, Blake published a book on the effects of government expenditure, which is more interesting for a modern reader than his 1810 work. He recanted his previous positions on depreciation, and maintained that during the inflation phase of the restriction of cash payments it was not paper to have depreciated, but gold to have risen. His argument was based on the importance of the foreign expenditure of the British government during the war years (a point largely neglected by the bullionists, and particularly by Ricardo). To this expenditure Blake attributed the fall in the exchange during the restriction, and saw the rise in the price of gold as a consequence of this fall (if the price of gold had not risen while the fall in the exchange continued, it would have been profitable to export it, but 'the holder of gold will not part with it, and transfer the power of making the profit to another person, unless at an advance in its price':1823, p. 15).

He explained the rise in the prices of commodities (other than gold) during the restriction, with the increased *internal* expenditure of the government. In the course of this second argument, he made interesting remarks on government expenditure, and criticisms of the orthodox positions. He rejected the arguments of those who maintained that government expenditure is only a transfer of demand from one channel to another, and that it would be 'derived from a fund that would have been equally a source of demand if it had been left

in the hands of the public' (1823, p. 44). He argued that the orthodox reasoning could not account for the great prosperity which during the war had accompanied the enormously increased government expenditure, and maintained that it was only '[i]f . . . the productive powers of the country were exerted to the utmost, and there was no means of adding to the gross annual produce', that government expenditure would be made 'at the expense of that fund which has before supplied the capitalist' (pp. 48–9). '[T]he error [of the contrary position] lies in supposing, first, that the whole capital of the country is fully occupied; and, secondly, that there is immediate employment for successive accumulations of capital as it accrues from saving' (p. 54).

Blake's 1823 book caused quite a stir in the orthodox camp. Ricardo intended to write a review of it, but he could not complete it before his death (the unfinished draft and extensive notes on the book, together with Blake's replies, have reached us: see Ricardo, *Works*, IV, pp. 325 ff.). Unfavourable reviews were published by McCulloch in *The Scotsman*, and by the young J.S. Mill in the *Westminster Review*. Malthus, on the other hand, declared himself largely in favour of Blake Malthus 1823, p. 72).

A distinguished economist in his own times (he also published a work, in 1839, on the assessment of tithes), he was afterwards almost entirely forgotten. A very short entry (by F.Y. Edgeworth) was devoted to him in Palgrave's *Dictionary*. Viner dismissed him as 'hopelessly confused' (1937, p. 203n), and he is not even mentioned in Schumpeter's *History of Economic Analysis*. Some attention to him is given by Corry (1958, pp. 41–5; see also Tucker 1960, p. 175).

Selected Works

1810. *Observations on the principles which regulate the course of exchange; and on the present depreciated state of the currency.* As reprinted in *A select collection of scarce and valuable tracts and other publications, on paper currency and banking*, ed. J.R. McCulloch, privately printed, London, 1857.

1823. *Observations on the effects produced by the expenditure of government during the restriction of cash payments*. London: Murray.

Bibliography

Corry, B. 1958. The theory of economic effects of government expenditure in English classical political economy. *Economica* 25: 34–48.

Malthus, T.R. 1823. *The measure of value, stated and illustrated, with an application of it to the alterations in the value of the english currency since 1790*. London: Murray.

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

Tucker, G.S.L. 1960. *Progress and profits in British economic thought. 1650–1850.*. Cambridge: Cambridge University Press.

Viner, J. 1937. *Studies in the theory of international trade*. Reprinted. Clifton: Kelley, 1975.

Blanc, Louis Joseph Charles (1811–1882)

J. Wolff

For Blanc, the problem of France is that of poverty; it is widespread:

> If there were only exceptional, isolated cases of suffering to alleviate, charity might perhaps be enough. But the causes of suffering are as general as they are profound, and it is by the thousand that one counts those amongst us who are deprived of clothing, food and shelter.

This suffering has its origins in competition which, through its tendency towards monopoly, has created poverty. Blanc popularized the idea that competition is destructive: workers, competing against each other, lower their wages; manufacturers bankrupt themselves in their struggle against each other. Moreover, the machine, instead of helping workers, only forces large numbers of them into unemployment and further accentuates competition. In contemporary society

the liberal thesis is straightforward: increasing the production of goods without taking account of their distribution leaves the fate of the weak at the mercy of chance. Work, which for the majority of people is the basis of their existence, is thrown into disarray and competition becomes systematic destruction rather than economic freedom. It is in no one's interest to maintain such a system because it deceives everyone. Social revolution is the only answer, and it must be undertaken because the existing social order will not last long and because revolution can be accomplished in a peaceful, orderly fashion.

What should be done? As the workers left to their own devices, cannot free themselves from their suffering, it is the duty of the government to help them. It must create workers' production associations, something Buchez had already advocated in 1834. But whereas Buchez thought above all of artisan and small businesses, Blanc had large industries in mind. By forming associations and cooperatives, workers would eventually win control of the means of production.

The State, which accordingly has to be strong, will initiate this movement for reform. It will be the 'poor man's banker'. After all, what can a free, talented man do if he has no capital? There is no liberty without real equality between citizens. The rights won by the Revolution of 1789 have no real force and no power to become effective. The State must distribute credit and so make it possible to create the tools of manufacture. It must set up a loan scheme to establish social workshops (*ateliers sociaux*) in all the main branches of industry.

All workers who so wished could work in these workshops, providing they showed some guarantee of dedication and morality. The hierarchy in these workshops would be established in their first year by the State, thereafter by election. In principal, all wages would be equal. Profits would be divided into three parts, the first to look after the old, the sick and the infirm, the second going towards easing crises in other industries, because all industries should help each other, and the third being set aside to help the workshops to expand by buying tools and instruments.

This socialization of the means of production will be achieved gradually. Private industries will progressively disappear, given the technical and social superiority of the social workshops; in effect, the social workshop constitutes a mode of organization where all workers, without exception, are encouraged to produce quickly and efficiently. The capitalists will not be expelled from the system, but will merely give up of their own accord. Although they may well be able to charge interest on the capital they invest in the workshops, Blanc refuses them any right to the profits, which will go first and foremost to the workers.

Moreover, this greatly increased feeling of cooperation and community would doubtlessly spread beyond the workplace. For Blanc, the 'obvious efficiency and incontestable richness of communal life' will give rise to the voluntary association of needs and pleasures.

Blanc was influenced by Necker and Sismondi but, like Turgot and Condorcet, he believed in progress. He wanted to build a new future without breaking with the past. He was the inventor of what came to be known as state socialism. He made specific that which Saint-Simon, Sismondi and Pierre Leroux had only sketched out. Later, interventionism came to enjoy a great deal of success in France and Germany, where it was propagated by Lassalle. But at the time Blanc was writing, public opinion, dominated by the liberals, thought the State incapable of effective intervention. An example of this is the rejection in 1838 of a bill proposing that the State complete the still largely unfinished railway network, and the decision to entrust this task to private companies. Blanc wanted the State to take this on, and he passionately defended his argument in the newspaper *Le bon sens*.

During the 1848 Revolution Blanc was a member of the Provisional Government and presided over the so-called Luxembourg Commission, which created the national workshops. But these were a mere caricature of his project for social workshops; they did not play their role of manufacture and instruction, and they foundered after disagreements between the members of the Provisional Government. After June 1848 Blanc

left Paris and fled to London. He returned to Paris after the fall of the Second Empire in 1871, when he became a member of the National Assembly. He sat on the far Left and declared his opposition to the Commune.

Selected Works

1839. *L'organisation du travail*. Paris.
1841. *L'histoire de dix ans*. Paris.
1848. *Le droit au travail*. Paris.
1847–62. *L'histoire de la Révolution*. Paris

Bibliography

Leroy, M. 1946. *Histoire des idées sociales en France*. Paris: Gallimard.
Loubere, L.A. 1961. *L. Blanc. His life and his contribution to the rise of French Jacobin Socialism*. Evanston: Northwestern University Press.

Blanqui, Jérôme-Adolphe (1798–1854)

Peter Groenewegen

Keywords
Blanqui, J.-A.; Economic history; History of economic thought; Say, J.-B

JEL Classifications
B31

French labour economist, economic historian and first major historian of economic thought, Blanqui was born in Nice and educated both there and in Paris, subsequently teaching humanities at the Institution Massin. His teaching brought him into contact with J.B. Say, who 'wished him for a disciple' (Blanqui 1880, p. ix) and to whose chair of political and industrial economy at the Conservatoire des Arts and des Métiers he succeeded in 1833. In addition, he was head of the Ecole Speciale du Commerce from 1830 to 1854, first editor of the *Journal des économistes* and from 1846 to 1848 served as member for Bordeaux in the Chamber of Deputies. In 1838 he was elected to the Académie des Sciences Morales et Politiques. He died in 1854 in Paris, more than a quarter of a century before his notorious younger brother, Louis Auguste, the revolutionary and member of the Paris Commune, with whom he is often confused.

Blanqui was a prolific writer but is now mainly remembered for his *Histoire de l'économie politique en Europe* (1837) which went through five editions. This is generally regarded as the first major history of political economy. In addition to doctrinal history it covered an enormous amount of economic history from the ancient world to the early 1840s. McCulloch (1845, p. 25) states that Blanqui's ancient economic history is 'brief and superficial; but his accounts of the political economy of the middle ages and modern times are more carefully elaborated, interesting and valuable. 'Blanqui's treatment of history reflects his support of free trade and sympathy for the working class. Schumpeter (1954, p. 498, n.18) praises Blanqui's 1826 *Resumé de l'histoire du commerce et de l'industrie* as a valuable historical monograph, while his *Précis élémentaire d'économie politique* is also worthy of notice.

Selected Works

1837. Blanqui, J.A. 1880. *Histoire de l'économie politique en Europe depuis les anciens jusqu 'â nos jours.* Trans. E.J. Leonard. *History of Political Economy in Europe*. New York: G.P. Putnam's sons.

Bibliography

McCulloch, J.R. 1845. *The literature of political economy.* London: LSE Reprint, 1938.
Schumpeter, J.A. 1954. *History of economic analysis.* London: Allen & Unwin, 1959.

Bloch, Marc (1886–1944)

R. Forster

In 1929 Lucien Febvre and Marc Bloch founded the *Annales d'histoire économique et sociale* (now called simply the *Annales*), a review that launched a new school of French historiography. Bloch and Febvre established this journal of sociological history as a '*une arme de combat*' against the traditional political and diplomatic history as taught by Langlois and Seignobos at the Sorbonne. Bloch and Febvre, colleagues at the University of Strasbourg since 1920, formed an ideal intellectual partnership. Febvre was arguably the more imaginative of the two superb scholars. A pioneer in what we now call the history of 'mentalities' and popular culture, Febvre drew upon cultural anthropology in his work a full generation before the 'Annales School' made this discipline one of its closest allies (*Le Problème de l'incroyance au XVIe siéle*: *la religion de Rabelais*, 1942). Bloch was, above all, a historian of Western agrarian regimes, meticulously explored over a millenium. His work reflects a thorough grounding in all of the historian's tools – archival, linguistic, geographic, archaeological and visual. Bloch was a medievalist by early training and his first work, *Les Rois thaumaturges* (1924) – a history of mentalities in its own right – gave little hint of his developing interest in economic history, a branch of history which had attracted little interest among French historians before Bloch was appointed to the Sorbonne in 1936. Bloch soon created an institute of economic and social history and planned a multiple-volume economic history of Europe, unfortunately never completed, except for his own *Esquisse d'une histoire monétaire de l'Europe* (1954).

Marc Bloch's most impressive achievement was his *Les Caractères originaux de l'histoire rurale Française* (1931), translated into English as *French Rural History*. In this now classic work on French rural society, Bloch traced a thousand years of history, demonstrating the slow evolution of field systems, farm technologies, the peasant household, the village community, communal usages, the incursion into the countryside of Church, State, noble and merchant, and the effects of long–run inflation on the various 'classes' in a complex rural hierarchy. One of Bloch's most original contributions was his linkage of a plough type (the heavy-wheeled plough drawn by heavy oxen or horses) to open elongated fields, which in turn necessitated communal farming and a whole package of collective rights of use. Bloch made a contrasting linkage between the light swing-plough, drawn by light oxen, to the closed irregular fields which necessitated fewer communal usages and led to a more absolute conception of private property. Yet in all of his hypotheses and interconnections, Bloch was extremely modest, always warning the reader of the limits beyond which the sources could not go.

Although his work pre-dates a more recent Annaliste awareness of ethnography and cultural anthropology, Bloch was a human geographer with a keen, even a visual grasp of *milieu* and locale. Appreciative of the work of folklorists such as Van Gennep, Bloch nevertheless preferred to identify the peculiar features of a locale by comparisons, among regions within France to be sure, but also with manors and fields on the other side of the Channel and the Rhine. His initial approach was rather to scrutinize a land survey (*cadastre*) or an aerial photograph of a field system than to 'decode' a village *fête*. At bottom, the 'original character' of French rural society was described by Bloch in structural rather than cognitive terms. Neither symbolic anthropologist nor econometrician, Marc Bloch was a positivist social historian who did not shy away from labels like 'agrarian individualism' when he thought them appropriate.

Marc Bloch's judiciousness, *bons sens*, and fair-mindedness, combined with his profound knowledge of every aspect of the agrarian structure from the 'gleaners of the stubble' to the precise curvature of the moldboard, has created great confidence in his work. This has been further reinforced by his personal testimony about history, especially in his *Métier d'historien* (1949), translated into English as *The Historian's Craft*.

Like many of his compatriots, especially among those who have also mastered agrarian history like Georges Lefebvre, Georges Duby or Emmanual LeRoy Ladurie, Marc Bloch was a model craftsman, not untouched by an underlying passion for the countryside.

Selected Works

1924. *Les rois thaumaturges*. Paris/Strasbourg: Faculté des lettres de l'Université de Strasbourg; London/New York: H. Milford, Oxford University Press. Trans. J.E. Anderson as *The royal touch*. London: Routledge & Kegan Paul, 1973.
1931. *Les caractères originaux de l'histoire rurale française*f. Paris/Cambridge, MA.: Harvard University Press. Trans. Janet Sondheimer as *French rural history: An essay on its basic characteristics*. Berkeley: University of California Press, 1966.
1939. *La Société fëodale*. Paris: A. Michel. Trans. L.A. Manyon as *Feudal society*. London: Routledge & Kegan Paul; Chicago: University of Chicago Press, 1961, 2 vols.
1941. *Testament*. Paris.
1946. *L'étrange défaite*. Paris: Société des Éditions Franc-tireur. Trans. G. Hopkins as *A strange defeat: A statement of evidence written in 1940*. London/New York: Oxford University Press, 1949.
1949. *Apologie pour l'histoire ou Métier d'historien*. Paris: Libraire Armand Colin. Trans. Peter Putnam as *The historian's craft*. New York: Vintage Books, 1964.
1954. *Esquisse d'une histoire monétaire de l'Europe*. Paris: A. Colin.
1958. *La France sous les derniers Capétiens, 1223–1378*. Paris: A. Colin.
1960. *Seigneurie française et manoir anglais*. Paris: A. Colin.
1969. *Souvenirs de guerre, 1914–1915*. Paris: A. Colin.
For the principal articles of Marc Bloch, see his *Mélanges historiques* (Paris: SEVPEN, 1963), vols 1–2. See also: Hommages à Marc Bloch, *Annales d'histoire sociale* VII and VIII, 1945.

Bodin, Jean (1530–1596)

D. P. O'Brien

B

Keywords
Bodin J.; Inflation; Just price; Monetary economics, history of; Quantity theory of money; Salamanca School; Scholastic economics

JEL Classifications
B31

Jean Bodin was born at Anger, France, in 1530 and died of plague at Laôn in 1596.

Bodin is chiefly famous to a wider public for works in history and philosophy. His first work to attract widespread attention has become known in English as *Method for the Easy Comprehension of History* (1566). But his *Republic* (1576), which deals with sovereignty as well as social justice (including proportional taxation), is generally regarded as his masterpiece. However, it is Bodin's work on inflation which is the most important part of his output for economists.

In developing this part of his work, Bodin had as background two key elements. The first was the 16th-century European inflation, triggered by imports of silver from the New World. Remarkable work by the American economic historian Earl J. Hamilton indicates something like a fourfold rise in prices in Spain during the 16th century (Hamilton 1934, 390–1, 493; see also Hauser 1932, xi–xix, xlvii–xlix). The Spanish inflation necessarily spread to Spain's immediate trading partner France, through official channels, informal ones (including smuggling), and piracy (Hauser 1932, xix–xxiv).

The second factor underlying Bodin's work was the contribution of Scholastic writers, stemming initially from an analysis of the effects of debasement, itself building upon the doctrine of the Just Price as founded on relative scarcity in a competitive market. If debasement of the

currency increased its nominal amount, its relative scarcity would decrease accordingly. A leading member of the School of Salamanca, Martín de Azpilcueta Navarro (1493–1586), applied this to money in general, whether debased or not, arguing that the purchasing power of money was inversely related to its quantity (Grice-Hutchinson 1952, 94–5).

Following Scholastic procedures, Bodin developed his own monetary analysis in the form of a critique of *Paradoxes* put forward by a writer called Malestroit. Malestroit's basic thesis was that, while prices had risen in terms of currency units as a result of debasement, they had not risen in terms of the precious metals. Utilizing data on changes in the price of land, Bodin's estimate of monetary inflation arising from depreciation of precious metals was in excess of 2.5 times, which is remarkably close to the level of 3.0 calculated by 20th-century economic historians.

Bodin's analysis of this inflation involved a treatment of the demand for money (he argued that this depended on the stage of economic development); of the importance of changes in the supply of money; of the idea that the money market clears; of disturbances to either demand for or supply of money producing price and/ or income changes; and of the direction of causality running clearly from monetary disturbances to the price level. All of these elements can be found in Bodin's response to Malestroit.

He had thus arrived at an important statement of the quantity theory. He did not claim that the fall in the value of silver was the sole cause of inflation; he certainly recognized the importance of debasement, and mentioned also monopolies, scarcity due to exports, and fashionable demand. But the increased supply of precious metals in France was of key importance.

Finally, Bodin recognized that inflation created economic uncertainty and interfered with economic activity. While changes in the supply of precious metals had to be treated as exogenous disturbances, inflation resulting from debasement should be checked, and he put forward a detailed case for currency reform.

See Also

- ▶ Inflation
- ▶ Just Price
- ▶ Monetary Economics, History of
- ▶ Scholastic Economics

Selected Works

1566. *Methodus, ad facilem historiarum cognitionem.* Paris: M. Iuuenem. Trans. B. Reynolds as *Method for the easy comprehension of history.* New York: Columbia University Press, 1945.

1568. *La Response de Maistre Jean Bodin Advocat en la cour au paradoxes de Monsieur de Malestroict, touchant l'enrichissement de toutes choses, & le moyen d'y remédier.* Paris: Martin le Jeune. Trans by H. Tudor and R. Dyson as *Responses to the paradoxes of Malestroit,* with an introduction by D. O'Brien. Bristol: Thoemmes, 1997.

1576. *Les Six Livres de la Republique de I. Bodin.* Paris: I. du Puys. Trans. R. Knolles as *Six books of a commonwealth.* London: Bishop, 1606.

Bibliography

Grice-Hutchinson, M. 1952. *The school of Salamanca: Readings in Spanish monetary theory 1544–1605*. Oxford: Clarendon.

Hamilton, E. 1934. *American treasure and the price revolution in Spain*. Cambridge, MA: Harvard University Press.

Hauser, H. 1932. *La Response de Jean Bodin*. Paris: Armand Colin.

O'Brien, D. 2000. Bodin's analysis of inflation. *History of Political Economy* 32: 267–292.

Böhm-Bawerk, Eugen von (1851–1914)

K. H. Hennings

Keywords

Austrian economics; Austrian school; Austrian theory of capital; Böhm-Bawerk, E. von; Capital deepening; Capital theory; Classical

economics; Distribution theory; Exploitation; Fisher I.; Goods theory; Imputation; Interest theory; Intertemporal exchange; Intertemporal preferences; Intertemporal theory of value; Inverse demand schedules; Knies, K. G. A.; Labour theory of value; Marginal productivity theory; Market power; Market rate of interest; Marxian economics; Menger, C.; Neoclassical economics; Price formation; Roundabout methods of production; Schäffle, A. E. F.; Shadow pricing; Stationary state; Subjective rates of interest; Subsistence fund; Time; Time preference; Wicksell effects; Wicksell, J. G. K.

JEL Classifications
B31

As civil servant and economic theorist, Böhm-Bawerk was one of the most influential economists of his generation. A leading member of the Austrian School, he was one of the main propagators of neoclassical economic theory and did much to help it attain its dominance over classical economic theory. His name is primarily associated with the Austrian theory of capital and a particular theory of interest. But his prime achievement is the formulation of an intertemporal theory of value which, when applied to an exchange economy with production using durable capital goods, yields a theory of capital, a theory of interest, and indeed a theory of distribution in which the time element plays a crucial role. Both this construction and his equally famous critique of Marx's economics strongly influenced the development of economic theory from the 1880s until well into the 1930s.

Eugen Böhm Ritter von Bawerk was born in Brünn (now Brno) in Moravia on 12 February 1851, the youngest son of a distinguished civil servant who had been ennobled for his part in quelling unrest in Galicia in 1848, and who died in 1856 as deputy governor and head of the Imperial Austrian administration in Moravia. After reading law at the University of Vienna, Böhm-Bawerk entered the prestigious fiscal administration in 1872. In 1875, however, after taking his doctorate in law, Böhm-Bawerk obtained a government grant to do graduate work abroad and prepare himself for a teaching position in economics at an Austrian university, as did his classmate and future brother-in-law Friedrich von Wieser. He worked for a year at Heidelberg with Karl Knies, and spent a term each at Leipzig, where Roscher taught, and at Jena, where Hildebrand taught. After working for another three years in the fiscal administration and the ministry of finance, he obtained his *Habilitation* (licence to teach) in 1880, and was immediately afterwards appointed to a professorship in economics at the University of Innsbruck, which he held until 1889. From a scholarly point of view, Böhm-Bawerk's years in Innsbruck were the most fruitful of his life. A book on the theory of goods, based on his *Habilitation* thesis, appeared in 1881, the first volume of *Kapital und Kapitalzins* in 1884. In 1886 he published a monograph on the theory of value in the most influential German language journal in economics, and in 1889 the second volume of *Kapital und Kapitalzins*. These publications established him as one of the leading members of the group of economists around Carl Menger who came to be known as the 'Austrian School'. In 1889 Böhm-Bawerk preferred an appointment in the Austrian ministry of finance to a chair at the University of Vienna because it carried the assignment to work out a reform of the Austrian income tax. He distinguished himself in the execution of this task, and rapidly rose in rank, obtaining the position of a permanent secretary in 1891, and in 1892 also the vice-presidency of a commission to assess the proposal of a return to the gold standard. Having been appointed minister of finance in a caretaker government in 1893, Böhm-Bawerk was considered to have risen too high to return to his former position when it was replaced by a parliamentary post after a few months, and he was made president of one of the three senates of the Verwaltungsgerichtshof, the highest court of appeal in administrative matters. In 1896 he was again made minister of finance in a caretaker government, but returned once more to the Verwaltungsgerichtshof in 1897. He was yet again appointed minister of finance in 1900, this time in a civil servants' government which fell

when he resigned in 1904 after large increases in military expenditure had been voted which he deemed threatened financial stability. This time he was offered, among other positions, the post of governor of the central bank, the most lucrative position in the monarchy. Yet he turned it down in favour of a chair at the University of Vienna which was especially created for him. Alongside Friedrich von Wieser (who had succeeded Menger in 1902) and Eugen von Philippovich, Böhm-Bawerk lectured on economic theory and conducted a seminar that soon attracted many able students, among them Joseph Schumpeter, Rudolf Hilferding, Otto Bauer, Ludwig von Mises, Emil Lederer and Richard von Strigl. He did not, however, return to the quiet life of a scholar. Having been elected a member of the Austrian Academy of Sciences in 1902, he was elected its vice president in 1907, and its president in 1911. He had also been made a Geheimrat (privy councillor) in 1895, had been appointed to a seat in the upper house of the Austrian parliament in 1899, and was from time to time given various other official assignments. Böhm-Bawerk died on 27 August 1914 at Rattenberg-Kramsach in Tyrol where he had tried to restore his health after having fallen ill on his way to a congress of the Carnegie Foundation in Switzerland as the official Austrian representative.

Böhm-Bawerk was as much a civil servant as a scholar, and in his later years an elder statesman in academic affairs as much as in the public realm of what was still a great power. He was extremely successful as an administrator and economic policymaker. But it is for his contributions to economic theory that he is chiefly remembered today. *Kapital und Kapitalzins* has become an economic classic even though it is defective in both construction and exposition. The first edition was written in great haste, and although Böhm-Bawerk responded over-conscientiously and meticulously to almost every criticism in the two further editions which appeared in his lifetime, adding so much material that two slim volumes grew into three massive tomes, he never found the time to rethink the structure as a whole. This absorptive attention to criticism was due to temperament as well as to circumstances. Böhm-Bawerk had a lawyer's mind and found it difficult to think in terms other than disjunct categories or 'cases' which needed to be distinguished sharply and did not fit into a continuum in which things shade into one another. Moreover, writing in a thoroughly anti-theoretical environment dominated by the German Historical School, he felt obliged to take issue and to sharpen differences for the sake of discussion. As a result, Böhm-Bawerk acquired an undeserved reputation as a casuistic and ungenerous controversialist which did much to place his (admittedly in some respects imperfect) contributions in a more critical light than they merit.

The core of Böhm-Bawerk's theoretical endeavours is the development of an intertemporal theory of value, capital and interest. This attempt owes much to his teachers in economics. A.E.F. Schäffle, Menger's predecessor in Vienna, seems to have convinced him that it was necessary to respond on a theoretical plane to the social question, the most pressing economic policy problem of the day, by developing a satisfactory theory of distribution (see Schäffle 1870). Karl Knies (1873–79) drew his attention to the problems of capital theory and the work of Marx. Carl Menger, finally, provided the starting point for his own theory.

In his *Grundsätze der Volkswirthschaftslehre* (1871), Menger had developed an atemporal theory of value, allocation and exchange. In his exposition and elaboration of that theory, Böhm-Bawerk (1886) strongly emphasized two of its aspects. Firstly, consumer behaviour is sharply distinguished from producer behaviour because only the former can evaluate goods directly; producers can do so only indirectly on the basis of their expectations of consumers' evaluations because production, being roundabout production, is necessarily time-consuming. Secondly, in both cases the evaluation of a commodity involves both the marginal utility of the commodity to the evaluating agent, and the marginal utility of the income available to him. In Böhm-Bawerk's usage, therefore, evaluations are shadow prices, or inverse demand schedules which imply an optimal allocation of commodities in the light of an agent's preferences as well as his income.

On the basis of such inverse demand schedules it was easy to show that the market price of a commodity could not be lower than the lowest price the 'last' buyer is prepared to offer, nor higher than the highest price the 'last' seller demands; here the 'last' seller is defined as the seller whose asking price is low enough to prevent any other seller from selling to the 'last' buyer: and the 'last' buyer as that buyer whose price offer is high enough to prevent any other buyer from buying from the 'last' seller. This definition, complicated as it is, is adapted to include the case of indivisible commodities which Böhm-Bawerk for one reason or another considered relevant.

Böhm-Bawerk also elaborated on Menger's seminal contribution by refining the analysis of distribution: he showed how inputs are evaluated by imputation, that is, by imputing to them their proper share of the value of the output they help to produce. In essence this amounted to a marginal productivity theory along lines laid down by J.H. von Thünen, but again adapted to his peculiarly Austrian assumptions of limited substitutability and finite divisibility of inputs.

Böhm-Bawerk generalized (in 1889) this theory of price formation in atemporal exchange to include intertemporal exchange by assuming that agents evaluate and trade not only currently available commodities, but also subjectively certain prospects of commodities available in the future. In his theory of goods, Böhm-Bawerk (1881) had shown in a surprisingly modern manner that such prospects exist, and how they can be evaluated. Assuming further that a market exists on which currently available commodities can be exchanged for subjectively certain prospects of commodities available in the future, the same argument can be applied to intertemporal exchange as was applied to atemporal exchange. Böhm-Bawerk did so in two stages, first considering a pure exchange economy without production, and then analysing an exchange economy with production.

In a pure exchange economy, all agents are consumers. Their inverse demand schedules, Böhm-Bawerk argued, involve for each agent a subjective rate of interest at which he is prepared, given his preferences over time and his (expected) income over time, to exchange subjectively certain prospects of commodities available in the future for the same amount of commodities available in the present. They also, Böhm-Bawerk maintained, typically exhibit positive time preference: commodities available in the present are typically evaluated at higher prices than subjectively certain prospects of the same commodities available in the future. This assertion is contained in the first two of three reasons he adduced for the positivity of the rate of interest. The first reason postulates that the marginal utility of income will decline over the planning horizon because of higher expected incomes in the future. The second reason postulates that for psychological reasons such as the finiteness of life, the marginal utility of a commodity declines as a rule with the length of time that elapses before it becomes available. As both these postulates have been much disputed it should be added immediately that Böhm-Bawerk regarded them as no more than testable assumptions which he deemed realistic but which admit exceptions. If these postulates are granted for all agents, their subjective rates of interest will always be positive, so that the market rate of interest will always be positive. The same will hold true if only the majority of agents behave according to these postulates. Böhm-Bawerk admitted that not all agents will always behave as postulated by him: but argued that as an empirical regularity they almost always did, and that his theory was applicable also when they did not. All that follows in the latter case is that the rate of interest is not positive. Note, therefore, that Böhm-Bawerk's argument establishes at one and the same time the existence of a (market) rate of interest in a pure intertemporal exchange economy, and identifies as the determinants of its height the relative intensities of the demand for, and supply of, commodities in the present and in the future, as expressed in agents inverse demand schedules. Of course, these are commodity rates of interest which do not necessarily exhibit any particular term structure, nor uniformity across different types of commodities. Both these properties need the further assumption that intertemporal markets exist for all commodities, and that at least some agents are prepared to

engage in arbitrage operations (see Nuti 1974), Böhm-Bawerk did not explicitly make these assumptions, but he argued as if these properties were assured. Note also that Böhm-Bawerk conceived in this model of a pure exchange economy of the rate of interest as a property of an intertemporal price structure, and not as the specific price for something, be it abstinence, the productivity of money, waiting, or whatever.

In order to extend the model just considered to include production Böhm-Bawerk argued that producers can be shown to have intertemporal inverse demand schedules like consumers, and postulated in his third reason that producers under-evaluate commodities available in the future on technical grounds. These assertions he derived from his analysis of the nature of production, and the role of capital in it. Production is assumed to be roundabout. It transforms non-produced or 'original' factors of production into consumable output with the help of capital goods which are internal to the production process. Because some capital goods are durable, production takes time. Böhm-Bawerk emphasized strongly the heterogeneity and specificity of capital goods. He also denied that they can be aggregated into some physical measure for the capital stock; aggregation is in his view possible only by valuing capital goods. He employed a forward-looking measure of capital value in which durable capital goods are valued by the present value of their services, and indeed generalized this procedure to all durable goods by showing that their valuation involves a subjective rate of interest which is equalized when durable goods are traded on markets.

The view of production as roundabout led Böhm-Bawerk to postulate a correspondence between the amounts of different capital goods used in production and the time which elapses before a particular dose of non-produced inputs has matured in the form of consumable output. This correspondence he formalized in the concept of a period of production which is defined as the average period for which the various doses of non-produced inputs required for the production of a unit output remain 'locked up' in the production process. This definition was a mistake which got him into more than one difficulty, and provided material for heated debates. To get round all the difficulties raised in these debates, assume that it is possible to define a period of production as a technical property of a particular production system which does not depend on factor prices; and assume further (with Böhm-Bawerk) that it can be used to order different methods of production in such a way that methods with a longer period of production can be said to be more capital intensive. More specifically, assume a temporal production function which (for a unit output) has only the period of production as argument, and which exhibits diminishing returns but is not homogeneous.

On this basis Böhm-Bawerk formulated a theory of producer behaviour in which competition forces producers to choose production methods that generate just enough output to pay the costs of production. As Böhm-Bawerk showed, this implied a discounted marginal productivity doctrine of (original) factor pricing, and hence the existence of positive quasi-rents at the margin. He also showed that this construction involved inverse demand schedules for capital goods which for each period of production define a profit maximizing rate of interest for given factor prices. At this point in his analysis, Böhm-Bawerk assumed the capital stock of an economy as given, and argued that the profit maximizing rate of profit can be determined with the help of that assumption. While that is correct it was another mistake which was duly seized upon (see for example Garegnani 1960) and which led to many debates. For the value of the capital stock associated with any method of production is an endogenous variable in his construction, as Böhm-Bawerk realized in other contexts. Nor was it necessary to make this assumption. It is sufficient to note that a single producer is forced by competition to pay neither less nor more than the discounted marginal value for the inputs he uses, if a time-consuming roundabout method of production is in operation. Translated into output prices this implies that he under-evaluates output available in the future. This is what Böhm-Bawerk asserted in the third reason; the technical ground being the method of production in operation. Note

that this is not so much a postulate or empirical regularity as it is an equilibrium condition.

Having thus established that producer behaviour can be characterized by derived inverse demand schedules for output which involve positive time preference, Böhm-Bawerk goes on to determine the market rate of interest in what is in effect a macroeconomic general equilibrium model. Attention is centred on the market for output available in the present, and the markets for claims to output available in the future. Supply on the market for output available in the present is fixed by decisions taken in the past; so is the supply available at all future dates whose production has already begun. Demand for output available in the present comes from consumers but will not exhaust supply if they save. Part of these savings will be taken up by other consumers in exchange for claims to output available in the future; transactions are consumption loans, and are likely, on Böhm-Bawerk's assumptions, to imply a positive rate of interest. Another part of savings will be taken up by producers, again in exchange for claims of future output, who use it to bid for more non-produced inputs in an attempt to expand the scale of production. As Böhm-Bawerk assumed that the amount of non-produced original factors is fixed, this results in higher factors prices and a change in the method of production (because higher factor prices can only be sustained if more output is produced). Net savings in the form of loans for productive purposes therefore imply a change in the method of production which, on Böhm-Bawerk's assumptions, implies capital deepening. Both kinds of transactions together determine the market rate of interest, which is thus seen to be determined by intertemporal consumer behaviour as summarized in the notion of positive time preference, and based on intertemporal preferences and the (expected) intertemporal distribution of incomes, on the one hand; and intertemporal producer behaviour as summarized in the period of production and the marginal product of extending it, and based on the intertemporal structure of roundabout methods of production on the other hand. Or, as Böhm-Bawerk put it, the rate of interest is determined by the relative evaluation of (output available in) the present and the future on the part of both consumers and producers. On his assumptions, this rate of interest is positive.

In some passages Böhm-Bawerk suggested that the rate of interest determined in his model is equal to the marginal product of an extension of the period of production. That created the impression that he had done no more than to establish, in a more roundabout way, what Jevons (1871, ch. 7) had already demonstrated. In other passages, however, Böhm-Bawerk seems to be aware that a change in the method of production involves a change in the value of the capital goods it requires, and that these Wicksell (or revaluation) effects imply that the rate of interest is less than the marginal product of an extension of the period of production. Böhm-Bawerk also obscured his argument by introducing the concept of a subsistence fund, thereby suggesting that his theory was no more than a revamped wages fund theory. Neither these nor other infelicities in his exposition should obscure the fact, however, that the hard core of his argument is the determination of the rate of interest as the property of an intertemporal price structure which in turn is determined by an intertemporal theory of value and allocation in consumption and production.

Böhm-Bawerk's model consciously referred to a stationary state as he wished to show that the rate of interest has something to do with the efficient allocation of resources in stationary as well as in non-stationary states. This comes out most clearly when he considers a socialist economy and demonstrates that it would require a positive rate of interest as does a capitalist economy. He did, however, consider non-stationary states in an interesting comparative static analysis of the effects of an increase in savings, and of technical progress. That he obtained a positive rate of interest in a stationary state is of course due to his assumptions, and no contradiction to Schumpeter's argument (1912) which is based on a somewhat different model (see Böhm-Bawerk, 1913, for a discussion of these differences).

The argument sketched on the preceding pages is expounded in Böhm-Bawerk's *Positive Theory* (1889) which he prefaced by a 'History and

Critique of Interest Theories' (1884) in which he critically examined earlier (and in later editions also contemporary) attempts to explain the rate of interest. The purpose of this volume has often been misunderstood. It is not a history of the subject which generously corrects mistakes, nor an attempt to differentiate his own product. Rather it is a 'negative theory' (Edgeworth): an attempt to survey the building blocks for his own theory and to pinpoint the pitfalls a satisfactory theory should avoid. Yet it cannot be denied that it is often overcritical. Thus Böhm-Bawerk shows again and again that the rate of interest cannot be said to be *determined by* marginal productivity considerations, but does not add that these nevertheless have a role to play in a more complete explanation. A similar omission occurs when he discusses abstinence or more generally intertemporal preferences.

One of the conclusions Böhm-Bawerk drew from his demonstration is that the existence of the rate of interest is not due to exploitation. It is obvious that on his argument workers can get the whole product of labour only if production is instantaneous. As long as production is roundabout, the present value of the workers' share in the value of the output they have helped to produce is necessarily less than what it would be if production were instantaneous. This is due, of course, to the existence of capital; but Böhm-Bawerk argued that interest would have to be paid irrespective of who owns such capital goods. That was also the gist of his critique of Marx's economics (1896), in which he singled out the labour theory of value as the basis of all errors. Böhm-Bawerk was (apart from Schäffle and Knies) one of the first economists to discuss Marx's economics on a scholarly plane; but he remained curiously blind to Marx's critique of the social institutions of a capitalist society. Although his critique drew a long reply from one of his students (Hilferding 1904) it was very influential and remained the best analytical performance of its kind until well into the 1950s (see Sweezy 1949).

Böhm-Bawerk's single-minded concentration on economic phenomena is also evident in his discussion of the role of economic power on markets (1914): in the short run, he argued, economic power may cause deviations from the state of affairs as defined by economic forces; in the long run, however, the latter will prevail. Again he was blind to any changes economic power may cause to the environment in which economic forces operate.

The impact of Böhm-Bawerk's work was immense, but its reception was made difficult by its prolixity and its technical defects, which offered many openings to critics. In essence, Böhm-Bawerk combined elements of neoclassical economic theory with elements of classical economic theory. He was neoclassical in his concern with rational economic behaviour and its consequences for the demand and supply of commodities, their pricing on markets, the forces which bring about equilibrium on markets, and the interaction of different markets. By contrast, classical lines of thought predominate in Böhm-Bawerk's analysis of production. However much he denied any adherence to classical cost theories of value, his view of production and the role of capital and time in it bear the mark of the Ricardian tradition.

The neoclassical part of his argument, in particular his analysis of intertemporal consumer behaviour, was taken up by Irving Fisher (1907, 1930) and developed into a theory of interest which is based on the notion of time preference (which Fisher transformed into a property of utility functions) and the concept of investment opportunities; these Fisher assumed rather than derived, thus cutting away Böhm-Bawerk's analysis of production and the role of capital in it. In this form, which admittedly offers insights into the problem of intertemporal allocation Böhm-Bawerk did not offer, Böhm-Bawerk's intertemporal theory of exchange became part of the heritage of orthodox neoclassical economic theory.

The more classical part of Böhm-Bawerk's model was taken up and elaborated by Wicksell (1893, 1901). In an attempt to free it of its classical garb, Wicksell turned it into a marginal productivity theory of the rate of interest. He ran into difficulties, however, not only over the proper definition of the period of production, but also because his neglect of what Böhm-Bawerk had to say about intertemporal consumer behaviour

forced him to assume a given capital stock in order to close his model. Wicksell used what had by then become the standard neoclassical concept of capital as a value sum, as proposed by J.B. Clark (1899), and (with good reason) combatted by Böhm-Bawerk. The shortcomings of such an argument, which was before long imputed to Böhm-Bawerk himself, were soon pointed out (see Cassel 1903; Garegnani 1960). Nevertheless Wicksell's interpretation became the standard portrayal of the 'Austrian' theory of capital and interest (see for example Lutz 1956; Dorfman 1959a, b; Hirshleifer 1967).

In the 1930s various attempts were made to reformulate Böhm-Bawerk's theory in such a way that it could be used as the basis of a theory of the short-run behaviour of an economy, particularly by Hayek (1931, 1939 and see Hicks 1967), but also by Hicks (1939, parts III and IV). This led to an intensive debate in which especially the capital theoretic foundations of his argument were examined, and found wanting (see Kaldor 1937; Reetz 1971, for a survey). There were some attempts at reconstruction (Eucken 1934; von Strigl 1934), but the definition of the period of production provided a major stumbling block. At the same time, Hayek and Knight repeated the debate between Böhm-Bawerk and Clark about the concept of capital on a somewhat different level. Finally Hayek (1941) made a major attempt to get round the difficulties the debate had shown up, and achieved some advances: but in the end his contribution turned out to be the final word that did not persuade anybody. The major difficulty which he did not manage to overcome was the fact that Böhm-Bawerk's construction does not lend itself to dynamic analysis precisely because his classical, macroeconomic approach to production and the role of capital requires an equilibrium approach, and does not provide a suitable basis for a discussion of producer behaviour out of equilibrium, and its dynamics.

More recent restatements of Böhm-Bawerk's argument consequently emphasize its static nature (von Weizsäcker 1971; Faber 1979), but do not really go beyond an exact formulation, in terms of modern capital theory, of some aspects of his theory. By contrast, Hicks (1973) is an innovative attempt to salvage some of the salient features of Böhm-Bawerk's view of production and capital, especially his emphasis on the role of time in production processes, in a modern framework which once more attempts to formulate a dynamic analysis (see also Belloc 1980, or Magnan de Bornier 1980). It centres on the concept of a 'transition' from one steady state to another, that is, a more long-term kind of economic dynamics than was considered in the 1930s; this is a promising approach which proves the vitality of Böhm-Bawerk's ideas.

Böhm-Bawerk posed a problem which had not been seen before in its full importance: the role of the rate of interest in the choice of an optimal method of production when production is roundabout, and its determination in a theory which takes seriously the impossibility of aggregating capital goods in physical terms. The solution he proposed is not without problems. But however much economic theory has progressed, some parts of his argument stand out as landmarks in the development of economic thought. Among them are his discussion of price formation on markets, especially those on which indivisible or finitely divisible commodities are traded, his analysis of time preferences, his analysis of intertemporal exchange, and his demonstration that the rate of interest is no more than a property of intertemporal price structures. His definition of the period of production turned out to be a cul-de-sac, but the possibilities his analysis of the role of time in production offers do not yet seem to have been exhausted.

Finally, the importance of his emphasis on the value aspect of the notion of aggregate capital and its implications has only recently been recognized as a seminal contribution. He can perhaps no longer be accorded the stature of a Ricardo or Marx. But the vitality of his ideas still ranks him among the great economists.

See Also

- ▶ Austrian Economics
- ▶ Period of Production

Selected Works

1881. *Rechte und Verhältnisse vom Standpunkte der volkswirtschaftlichen Güterlehre.* Innsbruck: Wagner. Trans. as 'Whether Legal Rights and Relationships are Economic Goods' in Böhm-Bawerk (1962).

1884. *Kapital und Kapitalzins. Erste Abteilung*: *Geschichte und Kritik der Kapitalzins-Theorien.* Innsbruck: Wagner. 2nd ed., 1900; 3rd ed., 1914; 4th ed., Jena: Fischer, 1921. Translation of 1st ed. as *Capital and interest.* London: Macmillan, 1890. Translation of 4th ed. as *Capital and interest.* vol. 1. South Holland: Libertarian Press, 1959.

1886. Grundzüge der Theorie des wirtschaftlichen Güterwerthes. *Jahrbücher für National-ökonomie und Statistik* 13, 1–82 and 477–541. Reprinted separately, London: London School of Economics, 1932.

1889. *Kapital und Kapitalzins. Zweite Abteilung*: *Positive Theorie des Kapitales.* Innsbruck: Wagner. 2nd ed., 1902; 3rd ed. in two volumes, 1909 and 1912; 4th ed. in two volumes, 1921, Jena: Fischer. Translation of 1st edn as *The positive theory of capital.* London: Macmillan, 1891. Translation of 4th edn as *Capital and interest,* vols. 2 and 3. South Holland: Libertarian Press.

1896. Zum Abschluss des Marxschen Systems. In *Staatswissenschaftliche Arbeiten, Festgaben für Karl Knies,* ed. O. von Boenigk. Berlin: Haering. Trans. as *Karl Marx and the Close of his System,* London: Fisher Unwin, 1898. Reprinted in Sweezy (1949). Also trans. as 'Unresolved Contradictions in the Marxian Economic System' in Böhm-Bawerk (1962).

1913. Eine 'dynamische' Theorie des Kapitalzinses. *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 22, 520–85 and 640–57.

1914. *Macht oder ökonomisches Gesetz? Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung 23, 205–271.*

1924. *Gesammelte Schriften,* ed. F.X. Weisz. Vienna and Leipzig: Hölder-Pichler-Tempsky.

1926. *Kleinere Abhandlungen über Kaptial und Zins.* Vienna and Leipzig: Hölder-Pichler-Tempsky.

1962. *Shorter classics.* South Holland: Libertarian Press.

Bibliography

Belloc, B. 1980. *Croissance économique et adaption du capital productif*. Paris: Economica.

Cassel, G. 1903. *The nature and necessity of interest*. London: Macmillan.

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

Dorfman, R. 1959a. A graphical exposition of Böhm-Bawerk's interest theory. *Review of Economic Studies* 26: 153–158.

Dorfman, R. 1959b. Waiting and the period of production. *Quarterly Journal of Economics* 73: 351–372.

Eucken, W. 1934. *Kapitaltheoretische Untersuchungen*. 2nd ed. Tübingen: Mohr.

Faber, M. 1979. *Introduction to modern Austrian capital theory*. Berlin: Springer.

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

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

Garegnani, P. 1960. *Il capitale nelle teorie della distribuzione*. Milan: Giuffrè.

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

Hicks, J.R. 1967. The Hayek story. In *Critical essays in monetary theory*, ed. J.R. Hicks. Oxford: Clarendon Press.

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

Hilferding, R. 1904. Böhm-Bawerks Marx-Kritik. In *Marx Studien* 1, ed. M. Adler and R. Hilferding. Trans. as 'Böhm-Bawerk's Criticism of Marx' in Sweezy (1949).

Hirshleifer, J. 1967. A note on the Böhm-Bawerk/Wicksell theory of interest. *Review of Economic Studies* 34: 191–199.

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

Kaldor, N. 1937. Annual survey of economic theory: The recent controversies on the theory of capital. *Econometrica* 5: 201–233.

Knies, K. 1873–9. *Geld und Credit.* 3 vols. Berlin: Weidemann'sche Buchhandlung.

Kuenne, R.E. 1971. *Eugen von Böhm-Bawerk*. New York: Columbia University Press.

Lutz, F.A. 1956. *Zinstheorie. Tübingen*: Mohr. Trans. as *The Theory of Interest*, Dordrecht: Reidel, 1967.

Magnan de Bornier, J. 1980. *Economie de la traverse*. Paris: Economica.

Menger, C. 1871. *Grundsätze der Volkswirthschaftslehre*. Vienna: Braumüller. Trans. as *Principles of Economics*. Glencoe: Free Press, 1951.
Nuti, D.M. 1974. On the rates of return of investment. *Kyklos* 27: 345–369.
Reetz, N. 1971. *Produktionsfunktion und Produktionsperiode*. Göttingen: Schwartz.
Schäffle, A.E.F. 1870. *Kapitalismus und Sozialismus*. Tübingen: Laupp.
Schumpeter, J.A. 1912. *Theorie der wirtschaftlichen Entwicklung*. Leipzig: Duncker & Humblot. Trans. as *The theory of economic development*. Cambridge, MA: Harvard University Press, 1934.
Schumpeter, J.A. 1914. Das wissenschaftliche Lebenswerk Eugen von Böhm-Bawerks. *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 23:454–528. Trans. as ch. 6 in *Ten great economists*, ed. J.A. Schumpeter, London: Allen & Unwin, 1952.
Schumpeter, J.A. 1925. Eugen von Böhm-Bawerk. *Neue österreichische Biographie 1815–1918*, 2. Vienna: Amalthea.
Stigler, G.J. 1941. *Production and distribution theories*. New York: Macmillan.
Sweezy, P.M., ed. 1949. *Karl Marx and the close of his system by Eugen von Böhm-Bawerk and Böhm-Bawerk's criticism of Marx*. New York: Kelley.
von Hayek, F.A. 1931. *Preise und Produktion*. Vienna: Springer. Trans. as *Prices and production*, London: Routledge.
von Hayek, F.A. 1939. *Profits, interest and investment*. London: Routledge.
von Hayek, F.A. 1941. *The pure theory of capital*. London: Routledge.
von Strigl, R. 1934. *Kapital und Produktion*. Vienna: Springer.
von Weizsäcker, C.C. 1971. *Steady state capital theory*. Berlin: Springer.
Wicksell, K. 1893. *Über Wert Kapital und Rente nach neueren nationalökonomischen Theorien*. Jena: Fischer. Trans. as *Value, capital and rent*. London: Allen & Unwin, 1954.
Wicksell, K. 1901. *Föreläsningar i Nationalekonomi. Vol. I: Teoretisk Nationalekonomi*. Lund: Berlinska Boktryckeriet. Trans. as *Lectures on political economy*. vol. 1, General theory. London: Routledge, 1934.
Wicksell, K. 1911. Böhm-Bawerks kapitalteori och kritiken därav. *Ekonomisk Tidskrift* 13, 39–49 Trans. in K. Wicksell, *Selected papers on economic theory*. London: Allen & Unwin, 1958.
Wicksell, K. 1914. Lexis och Böhm-Bawerk. *Ekonomisk Tidskrift* 16: 294–300. 322–34.

Bibliographic Addendum

Hennings, K. *The Austrian theory of value and capital: Studies in the life and work of Eugen von Böhm-Bawerk*. Cheltenham: Edward Elgar, 1997, is a detailed account.

Boisguilbert, Pierre le Pesant, Sieur de (1645–1714)

Peter Groenewegen

B

Keywords

Boisguilbert, P.; Cantillon, R.; Circular flow; Colbert, J.-B.; Equilibrium trade models; Hoarding; Physiocracy; Proportional income tax; Tax farming; Underconsumptionism

JEL Classifications

B31

French economist and lawyer. Born at Rouen into a *noblesse de robe* family, Boisguilbert was educated at a Jesuit college in Rouen, the city where he spent most of his life and where he died in 1714. The famous Port Royal and the Paris law school trained him as an *avocat* but initially inspired a literary career. This produced translations from the Greek (Dion Cassius and Herodotus) and some historical novels, one of which, *Marie Stuart, Reyne d'Ecosse* (1675) went through three editions. Marriage to a rich heiress in 1677 allowed him to pursue profitable activities in trade and agriculture for several years and enter the magistrature of Normandy. Such experiences brought home to him the deteriorating French economic position and the need to reverse this through fiscal and economic reform. His first economic work, *Le détail de la France* (1695) reflects these concerns. For the remainder of his life he unsuccessfully pressed plans for fiscal reform on various finance ministers, ultimately republishing his ideas, including the new *Factum de la France,* in various collected editions from 1707 (a detailed biography and bibliography is in Boisguilbert, 1966).

Boisguilbert is largely remembered as a precursor of the Physiocrats and as the economist whom Marx (1859, p. 52) linked with Petty as marking the start of classical political economy. His influence

was undoubtedly more extensive: much of Cantillon's (1755) circular flow analysis appears inspired by his work; while Roberts (1935, pp. 273–320) argues for considerable similarity between his fundamental economic ideas and some of Adam Smith's. A wealth of embryonic tools and concepts can be found in his work and include:

> division of labour, circular flow, velocity of money, hoarding, confidence, the multiplier, and variability of employment, supply and demand, diminishing utility, elasticity of demand, natural and market price, price variability, price flexibility, cobweb price-model, cost of production, diminishing returns, labour supply curve, bargaining range, impulse propagation, economic equilibrium, optimum and suboptimum price structures, and competition. (Spengler 1984, p. 77)

Tax criteria and class analysis need to be added to this list.

Boisguilbert's economic analysis ascribes France's economic distress to agricultural ruin from Colbert's edict prohibiting corn exports; excessive taxation worsened by tax farming; and financiers' power transforming money from a servant of trade into its tyrant. Underlying this diagnosis are models of equilibrium trade demonstrating the interdependence of the 200 occupations and professions exchanging their products at prices proportioned to necessary costs of production including a just profit. Hence buying, as the essential counterpart of selling and consumption, stimulates production. Disruptions to consumption prevent prices from covering costs, thereby initiating a downward spiral which ends in economic stagnation. Three causes for such disruptions are identified: low agricultural prices which lower rent and hence landlords' consumption demand; second, concentration of money among rich financiers leading to hoarding; third, lower consumption potential from excessive taxation. Since the livelihood of the poor depends on the consumption of the rich, unemployment and misery follow.

Boisguilbert's remedy follows from his identification of these causes of underconsumption. Free trade and encouragement of agriculture lead to a 'proper' corn price, conducive to high rents and consumption spending. Tax reform achieved by introducing a general proportional income tax removes the problem of excessive taxation and eliminates hoarding and leakages from the circular flow because the abolition of tax farming ends concentrated financier power. Subsequent encouragement of consumption allows prosperity to return and creates wealth for both the state and its citizens. Basic model, diagnosis and remedy are present with varying degrees of sophistication in Boisguilbert's major works, including *Traité de la nature, culture, commerce et intérêts des grains* (1704a) and *Dissertation de la nature des richesses, de l'argent et des tributs* (1704b), to name those not so far mentioned.

Selected Works

1695. *Le détail de la France.* Reprinted in Boisguilbert (1966), 581–662.

1704a. *Traité de la nature, culture, commerce et intérêt des grains, tant par rapport au public, qu'à toutes les conditions d'un état.* Reprinted in Boisguilbert (1966), 827–878.

1704b. *Dissertation de la nature des richesses, de l'argent et des tributs, ou l'on découvre la fausse idée qui règne dans le monde è l'égard de ces trois articles.* Reprinted in Boisguilbert (1966), 973–1012.

1707. *Factum de la France.* Reprinted in Boisguilbert (1966), 879–956.

1966. *Pierre de Boisguilbert ou la Naissance de l'économie politique.* Paris: Institut National d'Etudes Demographiques.

Bibliography

Cantillon, R. 1755. *Essai sur la nature du commerce en général*. London: Fletcher Gyles. Reprinted with an English translation and other material by Henry Higgs, London, for the Royal Economic Society, 1931.

Marx, K. 1859. *Contribution to the critique of political economy*. London: Lawrence & Wishart, 1971.

Roberts, H. Van Dyke. 1935. *Boisguilbert, economist of the reign of Louis XIV*. New York: Columbia University Press.

Spengler, J.J. 1984. Boisguilbert's economic views vis-à-vis those of contemporary *réformateurs*. *History of Political Economy* 16: 69–88.

Bonar, James (1852–1941)

Murray Milgate and Alastair Levy

Born at Collace in Perthshire (Scotland) on 27 September 1852, Bonar managed to combine a life-long career as a civil servant with the study of the history of economic thought, where his work focused on Smith, Ricardo, and especially Malthus. Somewhat ironically, given its rather poor reception at the time, his *Philosophy and Political Economy* (1893) is the book by which he is now principally remembered. Like Adam Smith, after graduating from Glasgow University, Bonar went as Snell Exhibitioner to Balliol College, Oxford, taking a first in 1877 and 'rounding-off' his studies at Leipzig and Tübingen (Shirras 1941, p. 146). Afterwards he removed to the contrasting environment of the East End of London, lecturing there for three years as one of the pioneers of the University Extension Movement and founding an Adam Smith Club to promote the popular discussion of economic matters. In 1881 he joined the Civil Service, in which he remained until his retirement in 1919. From 1907 he was Deputy Master of the Royal Mint in Ottawa.

Bonar's early services to Ricardo scholarship were rendered in two compilations of correspondence: *Letters of David Ricardo to Thomas Robert Malthus*: *1810–1823*, and *Letters of David Ricardo to Hutches Trower and Others*: *1811–1823*, published in 1887 and 1899 respectively (the latter jointly with Jacob Hollander). In neither instance was Bonar able to recover the 'missing' Malthus and Trower letters which were not unearthed until the discovery of the so-called 'Ricardo Papers' in 1930. These are now published in Sraffa's edition of Ricardo's *Works and Correspondence*. In the Bonar edition of the letters to Malthus a number of errors of dating occurred, giving a rather misleading picture of the temporal development of Ricardo's work – most of these errors arose, it seems, from a mis-reading of Ricardo's handwriting, where the number '3' closely resembles a broken '0' (see Keynes 1933, p. 112 n.2; Sraffa 1951–73, vol. VI, p. xxi n.1).

To Smithiana, Bonar bequeathed *A Catalogue of Adam Smith's Library* (1894), a handsomely printed volume even by the standards of the day, and *The Tables Turned* (1926), an imaginary discussion whose participants included (in addition to Smith), Ricardo, Malthus, Mill and Marx. It is difficult to believe that Bonar could have anticipated the quite extraordinary bout of antiquarianism that infected Smith studies as a consequence of the appearance of his catalogue. Around the substantive question of the extent of the direct indebtedness of Smith to Physiocracy (and, particularly, to Turgot) – where the content of the library is one quite minor piece of evidence – there sprang up an industry designed to track down, it would seem, every last item. The bug infected not only Scottish writers, but also American and Japanese economists, much to the detriment of obtaining a satisfactory answer to the original question.

Bonar's enthusiasm for Malthus (his 'services to general theory are at least equal to Ricardo's' 1885, p. vii) was rivalled only by that of Keynes. Aside from two books (1881, 1885), his entry in Palgrave's *Dictionary*, and an *Economic Journal* article (1929), Bonar was engaged for much of his life on a full-scale intellectual biography of Malthus. How far this might have advanced the understanding of Malthus's contribution is impossible to say. However, since Bonar's published writings on Malthus appeared before both the discovery of Malthus's side of the Ricardo correspondence (in 1930) and Keynes's celebrated essay on Malthus (in 1933), the availability of additional material could hardly have failed to lead Bonar into new fields of interpretation.

While it was largely the above-mentioned works that secured Bonar's reputation during his lifetime – leading to honorary degrees from Glasgow and Cambridge, and election to the British Academy – much of it has now been superseded. His *Philosophy and Political Economy*, however, has proved more resistant to the passage of time. Its discussion of utilitarianism, for example, can

still be read with profit. In places, Bonar's command of things 'German' – from Kant, Fichte and Hegel down to Richard Wagner – is impressive. Furthermore, with an early article (1888–9) and an entry for Palgrave's *Dictionary of Political Economy*, Bonar is credited in some circles with introducing the work of the Austrian School to an English-speaking audience.

At the age of sixty, Bonar climbed the Wetterhorn (3,708 m) in a snowstorm – a feat which pales into insignificance when measured against the effort it must have required to complete upwards of seventy entries for the original edition of this Dictionary. He died on 18 January 1941 at the age of eighty-eight; his 'definitive' biography of Malthus (Keynes 1933, p. 81n), the manuscript of which Shirras claimed to 'have with him' in 1941 ready for post-war publication, remains unpublished.

Selected Works

1881. *Parson Malthus*. Glasgow: James Maclehose.
1885. *Malthus and his work*. London: Macmillan; 2nd edn, London: G. Allen & Unwin, 1924.
1887. *Letters of David Ricardo to Thomas Robert Malthus: 1810–1823*. Oxford: Clarendon Press.
1888–9. Austrian economists and their view of value. *Quarterly Journal of Economics* 3, October, 1–31.
1893. *Philosophy and political economy*. London: Swan Sonnenschein & Co; 4th edn, London: G. Allen & Unwin, 1927.
1894. *Catalogue of Adam Smith's library*. London: Macmillan; 2nd edn, 1932.
1899. (With J.H. Hollander.) *Letters of David Ricardo to Hutches Trower and others: 1811–1823*. Oxford: Clarendon Press.
1922. Knapp's theory of money. *Economic Journal* 32, March, 39–47.
1926a. Memories of F.Y. Edgeworth. *Economic Journal* 36, December, 647–53.
1926b. *The tables turned. A lecture and dialogue on Adam Smith and the classical economists*. London: P.S. King & Sons.
1929. Ricardo on Malthus. *Economic Journal* 39, June, 210–18.
1931. *Theories of population from Raleigh to Arthur Young*. London: G. Allen & Unwin.

Bibliography

Keynes, J.M. 1933. *Essays in biography*. London: Macmillan. In the *Collected works of John Maynard Keynes*, Vol. X, London: Macmillan, 1972.
Shirras, G.F. 1941. James Bonar. *Economic Journal* 51: 145–156.
Sraffa, P. (ed.) 1951–73. *The works and correspondence of David Ricardo*, 11 vols. Cambridge: Cambridge University Press.

Bondareva, Olga (1937–1991)

Myrna Wooders

Keywords
Balancedness; Approximate balancedness; Bondareva, O.; Clubs; Coalitions; Cooperative games; Core equivalence; Cores; Duality; Essential superadditivity; Game theory; Linear programming; Local public goods; Market games; Price-taking equilibrium

JEL Classifications
B31

Olga Nikolajevna Bondareva was born in St Petersburg on 27 April 1937. She joined the Mathematical Faculty of the Leningrad State University in 1954, and completed a Ph.D. in mathematics at the Leningrad State University in 1963, in part under the supervision of Nicolaj Vorobiev. Her thesis was entitled 'The Theory of the Core in an n-Person Game'. Bondareva rose through the ranks at Leningrad State University to become a senior research fellow in 1972 and a leading research fellow in 1989. Because she sympathized with a student who wished to emigrate to Israel,

however, she was prohibited from teaching from 1973 until 1989. With *perestroika* and increased freedom to travel outside the Soviet Union, she became an active and energetic international figure in game theory. She died as a result of a traffic accident on 9 December 1991.

Bondareva published over 70 works on game theory and mathematics, supervised seven Ph.D. students, and was a member of the editorial board of *Games and Economic Behavior*. Her work on the core of a cooperative game plays a central role in game theory, and her insights can be seen underlying recent work on the theory of price-taking equilibrium and the core.

The following is a brief description of Bondareva's celebrated result. To allow us to see the relationship of this result to more recent research on games and economies with many players, it is stated for games with player types and requiring only 'essential superadditivity' in the definition of feasible payoffs.

Define a (pre)game with T types of players as a function ψ from vectors of non-negative integers $s \in \mathbb{Z}_+^T$, $s \neq 0$, called *profiles* of coalitions, into the non-negative real numbers $\mathbb{R}_+$. Given a vector $m \in \mathbb{Z}_+^T$, representing the total player set of the game and $s \in \mathbb{Z}_+^T, s \leq m, \psi(s)$ is interpreted as the total payoff to a coalition of players consisting of s_t identical players of type t, $t = 1, \ldots, T$. Let $(s^\ell : \ell = 1, \ldots, L)$ denote the collection of all profiles $s^\ell \leq m$. A *partition of a profile s* is determined by a collection of non-negative integers $(n_1, \ldots, n_L)$ satisfying the condition that $\sum n_\ell s^\ell = s$. With the domain of ψ restricted to profiles $s^\ell \leq m$, the pair (m, ψ) determines a cooperative game. Let $\psi^*(m)$ denote the maximum, over all partitions of m, of $\sum n_\ell \psi(s^\ell)$. A payoff vector $\overline{x} \in \mathbb{R}^T$ is in the (equal treatment) *core* if and only it holds that. $\overline{x} \cdot m \leq \psi^*(m)$ ($\overline{x}$ is fesible) (x is *feasible*) and for each ℓ, $\psi(s^\ell) \leq \overline{x} \cdot s^\ell$.

Now consider the following linear programming (LP) problem:

$$\min_{\overline{x}} \overline{x} \cdot m \text{ subject to } \psi(s^\ell) \leq \overline{x} \cdot s^\ell \text{ for all profiles } s^\ell \leq m.$$

A vector $\overline{x}^*$ is in the core if it is a solution to the above LP problem *and* $\overline{x}^* \cdot m = \psi^*(m)$. The dual LP problem is:

$$\max_{\omega_1, \ldots, \omega_L} \sum_\ell \omega_\ell \psi(s^\ell) \text{ subject to } \sum_\ell \omega_\ell s^\ell = m \text{ and } \omega_\ell \geq 0 \text{ for all } \ell$$

From the fundamental duality theorem of linear programming, there is a solution to the first LP problem if and only if there is a solution to the second, and, in this case, it holds that the optimal values of the objective functions in the two LP problems are the same.

For the second LP problem, let (ω_ℓ^*) denote the solution for the 'balancing weights' (ω_ℓ). The game is *balanced* if and only if $\sum_\ell \omega_\ell^* \psi(s^\ell) = \psi^*(m)$. It follows that a game is balanced if and only if it has a non-empty core, Bondareva's result. (See also Shapley 1967.)

Numerous applications of game theory to economics have employed the concept of balancedness. An outstanding contribution is Shapley and Shubik (1969), who show an equivalence between the set of totally balanced games (balanced games with the property that every subgame also has a non-empty core) and market games (cooperative games derived from economies where all players have concave utility functions). Bondareva's result as formulated above is a key ingredient in Wooders (1994), showing that under mild conditions games with many players are market games. Scarf (1967) demonstrates non-emptiness of the core of a balanced game without side payments (where the payoff set for a coalition S is a subset of RS rather than a real number). Bondareva's result also underlies the approximate balancedness of economies with clubs or relatively small effective or nearly effective coalitions. While this result has been demonstrated in much generality, the key is simple. Since the coefficients of the dual LP problem are integers, when the total player set is replicated (becomes rm, $r = 1, 2, \ldots$) and no new effective coalitions are permitted (that is, if $\psi(s) \geq 0$ then $s \leq m$) then there is an integer k such that all replicated games with total player profiles given by rkm are balanced

(Wooders 1994, and references therein). The integer k clears the denominators of the (rational) extreme points of the convex set of balancing weight vectors of the dual LP problem. In recent works on the theory of clubs and local public goods, balancedness plays a crucial role; see Demange and Wooders (2005) for several recent examples and additional references.

We refer the reader to Rosenmueller (1992) for some additional details of Olga Bondareva's life. See also Kannai (1992) for an excellent review of research on the core and balancedness.

See Also

► Game Theory

Selected Works

1963. Some applications of linear programming to the theory of cooperative games. *Problemy Kibernetiki* 10, 119–39 [in Russian]. English translation in *Selected Russian Papers in Game Theory 1959–1965*. Princeton: Princeton University Press, 1968.

1969. Solution for a class of games with empty core. *Soviet Doklady* 185(2).

1975. Acyclic games. *Vestnik of Leningrad University* No. 7.

1978. Convergence of spaces with a relation and game-theoretical consequences. *Zhurnal Vych. Math. i Math, Phys*. 18(1).

1979. Development of game theoretical methods of optimization in cooperative games and their applications to multi criterial problems. In *Sovremennoe sostojanie terorii issledovaniji operatiskil* [State of the art in the theory of operations research], ed. N. Moiseev. Moscow: Nauka.

1983. Extensive coverings and some necessary conditions of existence of solutions in cooperative games. *Vestnik of Leningrad University* No. 19.

1987. Finite approximations of choice on infinite sets. *Izvestija AN SSSR, Tekhnicheskaja kibernetika* No. 1.

1989. Domination, core and solution (a short survey of Russian results). Discussion Paper No. 185. IMW, University of Bielefeld.

1990a. Game theoretical analysis of one product market with similar utilities. *Kibernetika* No. 5.

1990b. Revealed fuzzy preferences. In *Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory*, ed. J. Kacprzyk and M. Fedrizzi. Dordrecht: Kluwer Academic Publishers.

1994. (With T. Driessen.) Extensive coverings and exact core bounds. *Games and Economic Behavior* 6, 212–219.

Bibliography

Demange, G., and M. Wooders, eds. 2005. *Group formation in economics; Networks, clubs and coalitions*. Cambridge: Cambridge University Press.

Kannai, Y. 1992. The core and balancedness. In *Handbook of game theory with economic applications*, ed. R. Aumann and S. Hart. Amsterdam: North–Holland.

Rosenmueller, J. 1992. Olga Nikolajevna Bondareva: 1937–1991. *International Journal of Games Theory* 20: 309–312.

Scarf, H. 1967. The core of an n-person game. *Econometrica* 35: 50–69.

Shapley, L. 1967. On balanced sets and cores. *Navel Research Logistics Quarterly* 9: 45–48.

Shapley, L., and M. Shubik. 1969. On market games. *Journal of Economic Theory* 1: 9–25.

Wooders, M. 1994. Equivalence of games and markets. *Econometrica* 62: 1141–1160.

Bonds

Donald D. Hester

Abstract

A bond is commonly understood to be a debt instrument in which a borrower receives an advance of funds and contracts to make future payments of interest and principal according to an explicit schedule. The nominal return from holding a bond is the sum of its interest

payments and the change in its price over an arbitrary holding period. Bonds differ in terms of face value, maturity, callability, seniority, convertibility, risk of default, and size, frequency and taxability of interest payments. Since 1970 bond markets have experienced a number of major institutional changes with enduring consequences for capital markets.

Keywords

Asset-backed debt; Bankruptcy; Bonds; Capital gains and losses; Central banks; Default; Defeasance; Derivative securities; Discount bonds; Eurobonds; Interest rates; Junk bonds; Medium term notes; Miller, M.; Modigliani, F.; Options; Sovereign debt; Stripped bonds; Tobin, J.

JEL Classifications

G1

A bond is a contract in which an issuer undertakes to make payments to an owner or beneficiary when certain events or dates specified in the contract occur. The term has medieval origins in a system where an individual was bound over to another or to land. Subsequently, goods were put in a bonded warehouse until certain conditions (for example, payments of taxes or tariffs) were satisfied; individuals were released from jail when a bail bond guaranteeing their appearance in court was supplied; and individuals were allowed to perform certain tasks when a surety or performance bond guaranteeing satisfaction was provided. Governments and individuals have borrowed from others since earliest recorded history, as Sumerian documents attest. Perhaps public bonds first appeared in modern form with the establishment of the Monte in Florence in 1345. Monte shares were interest bearing, negotiable and funded by the Commune.

In contemporary economic discourse, a bond is commonly understood to be a debt instrument in which a borrower, typically a government or corporation, receives an advance of funds and contracts to make future payments of interest and principal according to an explicit schedule. The remainder of this entry focuses almost exclusively on these debt instruments. Terms of bonds are designed to protect the rights of borrowers and creditors; they are heterogeneous and their interpretations and enforceability vary across legal jurisdictions.

Bond Heterogeneity

The distinction between bonds and other evidences of debt such as loans or notes is inherently arbitrary and imprecise. Bonds tend to have long specified maturities when issued, or none at all in the case of consols. However, issuers may reserve the right to call them after they have been outstanding for a specified time interval. Other things being equal, bonds that are callable have higher rates of return than those with no call provision, because issuers have an incentive to call them whenever market rates fall below rates that existed when the bonds were offered. While bonds ordinarily convey no equity stake in an enterprise, some corporate bonds are convertible; they include a clause that gives bondholders an option to convert bonds to shares of the issuer's common stock at a specified conversion value in some time interval. Other things being equal, convertible bonds have lower interest rates than bonds with no conversion rights, because the option to convert is valuable. Formulas for determining the values of options are discussed by Black and Scholes (1973) and Zhang (1997).

Bonds tend to be negotiable and can usually be traded on an established secondary market. Once bonds are issued, bondholders are strategically vulnerable to actions of a firm's management, equity holders, and short-term lenders, as has been argued by Bulow and Shoven (1978), especially if an issuer's financial condition deteriorates. Default occurs if a bond issuer fails to make scheduled payments of interest or principal or violates other covenants of a contract. A bondholder's rights in a default situation are circumscribed by the terms of the contract and by judicial authority.

In the event of a default by a corporation, bondholders or other interested parties may petition for protection under bankruptcy statutes. In

some circumstances a bankruptcy court appoints a receiver to conserve the value of a firm's assets so as to protect creditors. The fraction of a creditor's claims that is paid is determined in part by their seniority (or priority) relative to other claims. Bonds may be either unsubordinated or subordinated to other debt. A bankrupt firm may be liquidated in favour of its creditors or be reorganized and allowed to continue with partial payouts to creditors.

In the United States, bonds issued by corporations and state and local governments are assigned credit ratings by firms such as Moody's Investors Services and Standard and Poor, Inc. Bonds with lower credit ratings are predicted to have a higher rate of default; they tend to have higher *ex ante* rates of return to compensate holders for higher expected default losses and risk of default. Bonds of state and local governments fall into two broad classes: (*a*) bonds which are general obligations of the issuing government and (*b*) revenue bonds, where interest and principal payments are dependent on income from some specific project. Because general obligation bonds are funded from taxes of the issuer, they tend to have higher ratings and lower rates of interest than revenue bonds. Corporate bonds with poor credit ratings are called 'junk bonds'. Before 1980, most bonds had been issued with good ratings and were suitable for the portfolio of a prudent investor. If an issuer's condition subsequently deteriorated, its bonds were downgraded and possibly became junk bonds. Beginning in about 1982, this practice changed and large amounts of funds were raised by issuing bonds that had low ratings when first offered. The reasons for offering junk bonds are incompletely understood but include avoidance of corporate income taxes, as was predicted by Modigliani and Miller (1963). Coinciding with the issuance of junk bonds were a substantial increase in leverage (the ratio of a firm's debt to net worth) and a wave of leveraged buyouts in which publicly traded corporations were reorganized into enterprises that were narrowly held by management and a few outside investors.

The significance of these changes in imperfect capital markets is controversial; in traditional financial theory it is often argued that high leverage makes a firm vulnerable to financial shocks and recessions. High leverage is believed to reduce the probability of a firm being taken over or bought up. Leverage on the books of a firm, however, can be misleading without knowledge of the contractual rate of interest on a firm's bonds. For example, when interest rates rise a firm may call its existing low-interest rate bonds which have a low market price and finance them with a smaller quantity of new bonds that bear the new high rates. This action, 'defeasance', reduces the ratio of debt to equity on a firm's books without reducing its interest costs.

Bonds issued by autonomous nation states are 'sovereign' debt. Defaults by issuers of sovereign debt do not result in bankruptcy proceedings, because there is no world bankruptcy court and applicable code. Moreover, as Bulow and Rogoff (1988) have argued, there is no credible basis for establishing seniority among sovereign debt issues in the event of a default. Sovereign bonds that default are traded at deep discounts for indefinitely long periods. While bankruptcy is impossible, negotiations leading to the restructuring of a country's debt obligations do occur, and sanctions against a defaulting country have been imposed by other countries where bondholders are concentrated. Credit ratings of sovereign debt vary widely across countries and, in part, are a function of the bond repayment history of a country.

Bond Yields and Rates of Return

The 'yield' on a bond is the flow of interest income to its holders. Apart from defaults, bonds traditionally pay interest in fixed amounts on specified dates that are indicated by coupons on the bond. Coupon-bearing bonds may allow investors to choose portfolios that match interest and amortization streams with their own nominal future requirements for funds. A portfolio is said to be perfectly 'immunized' against interest rate fluctuations if such matching is achieved. Bonds that have no coupons are called 'discount bonds'; they provide no interim cash flow and are retired at maturity with a payment equal to their face or par value, which is higher than the issue price.

Default-free discount bonds thus afford nominal *income certainty* to investors, as was explained by Robinson (1951), but do not guarantee that an investor's spending goals can be achieved when inflation is unpredictable. Some protection against inflation is afforded by inflation-indexed bonds that first appeared in Israel in 1955, the United Kingdom in 1981 and the United States in 1997, when US Treasury Inflation-Protected Securities (TIPS) were first offered. With TIPS, protection takes the form of a percentage increase of the bond's principal that equals the rate of inflation. Because the increase is taxable and inflation is based on the rate of change of the consumer price index, TIPS only incompletely protect a representative investor against inflation. For a discussion, see Wrase (1997).

The nominal return from holding a bond is the sum of its interest payments and the change in its price over an arbitrary holding period. For example, if there are no transactions costs and taxes, the return from holding a multi-year bond for two years is:

$$\text{return} = y_1 + y_2 - P_p + P_s \qquad (1)$$

where P_p and P_s are respectively the purchase and selling price and y_1 and y_2 are annual interest payments. If interest payments are assumed to be paid at year end, the nominal annual rate of return, r, from this two-year investment is obtained by solving the polynomial:

$$P_p = y_1/(1+r) + (y_2 + P_s)/(1+r)^2 \qquad (2)$$

If the bond is bought at P_p and sold at P_s, a bond trader is said to 'realize' a capital gain (loss) if P_p is less (more) than P_s.

A condition for equilibrium in a bond market is that expected rates of return from holding similar bonds are similar. If this condition were not satisfied, bond traders could improve portfolio earnings through arbitrage, by selling the bond with the lower rate of return and buying the bond with the higher rate of return, so long as the difference exceeds transactions costs. When transactions costs are zero, bonds are perfectly 'reversible'. When market rates of return rise, prices on outstanding bonds fall and rates of return experienced by existing bondholders fall; capital losses are sustained by holders of all but maturing bonds. Bond traders attempt to buy bonds immediately before market rates of return fall so that they may realize capital gains by buying at a low price and selling at a high price. Similarly, speculative traders of bonds seek to sell bonds immediately before market rates of return rise. While bonds that do not default mature at par, the prices of outstanding bonds are incompletely predictable; generally bonds with more years to maturity have more price volatility.

Bond Issuance Considerations

Bonds are issued by governments and corporations to finance deficits and acquire assets. While neither issuer can afford to ignore imminent movements in interest rates, their time schedules of outlays are somewhat inflexible. Deficits must be financed, and it is short-sighted to delay purchasing high rate-of-return assets to take advantage of transient interest rate movements. Firms needing funds may choose to finance a long-term asset with short-term borrowings from banks, with a long-term bond whose interest rate varies (or 'floats') over time in a fixed relation to short-term rates, or with a long-term fixed coupon bond. Bank borrowing to finance long-term assets exposes firms to the risk that banks may unilaterally alter loan terms or refuse to renew maturing loans. Firms avoid non-renewal risk by borrowing with bonds. A firm's choice between issuing conventional fixed-rate bonds and floating rate bonds to finance an asset depends in part on the correlation between returns from the asset being acquired and short-term interest rates for reasons that are developed by Cox et al. (1981). Other things being equal, a floating rate bond exposes a firm to less risk when the short-term rate and the rate of return on the acquired asset are positively correlated.

Government deficits are financed by issuing short-term bills, notes, bonds and 'outside' or fiat money. Central banks control the ratio of outside money to interest-bearing government

debt when conducting monetary policy. Central bank sales (purchases) of bonds decrease (increase) bond prices and increase (decrease) bond interest rates in the market. Other things being equal, an increase in bond interest rates increases the cost of financing new capital equipment and causes marginal investment projects to become unprofitable. Control of bond and other market interest rates by central banks is one handle through which monetary policy affects the level of macroeconomic activity. It has also been argued by Tobin (1963) that the composition of outstanding interest-bearing government debt can importantly influence the level of macroeconomic activity. If bonds are closer substitutes for physical capital in investors' portfolios than are treasury bills, a debt management policy of selling bonds and buying an equivalent amount of bills discourages private sector capital formation.

Recent Innovations in Bond Markets

Since 1970 capital markets have experienced a number of major institutional changes and innovations that have had enduring consequences for bond markets. Arguably the most important was the introduction of securitized debt by the US government-sponsored enterprises, Federal National Mortgage Association and Federal Home Loan Mortgage Corporation, and by the Government National Mortgage Association. While they could issue conventional bonds, they also could issue what amounts to second-order bonds, such as pass-through securities or collateralized mortgage obligations. Instead of an issuer being responsible for paying interest and retiring principal, securitized debt replaces the issuer with a constructed package of mortgage loans that generates a stream of interest and principal payments to holders of the securities. Initially, the underlying loans were insured against default, but they differed from traditional bonds because mortgage loans could be paid off before their contractual maturity. Thus, these securities were bonds with discrete, stochastic call provisions. The underlying stochastic process is in part a function of past and current market interest rates, because homeowners tend to refinance their houses when market interest rates fall.

In 1985, securitized debt evolved into generalized asset-backed debt, which serves to finance a package of self-liquidating financial assets. Like bonds, some of this debt is publicly rated for safety by investment services, but much of it is privately placed and not traded on a secondary market where ratings are important. The value of the assets underlying a debt issue typically exceeds the face value of the issue by an amount called a 'haircut', which serves as a partial safeguard against default. Asset-backed debt is heterogeneous; interest rates may be fixed or indexed to some market rate, amortization schedules vary, and the qualities of underlying assets differ. In 2004, new issues of asset-backed debt exceeded new issues of conventional bonds by corporations and governments for the first time. Asset-backed debt tends to be less costly to issue and to service, which largely accounts for its rapid growth. It is often issued by a 'special purpose vehicle', a legal entity which is intended to be bankruptcy-remote and whose sole function is to service a set of debt issues. Unlike conventional bonds, such debt usually does not appear on government or corporate balance sheets, which partly explains its appeal in a world where leverage has been rising. However, especially in Europe there is a hybrid 'covered bond', which is a securitized bond that remains an obligation of the issuer and continues on balance sheets. Because it is collateralized, it retains value even when the issuer fails.

Another innovation that has partly displaced bonds are medium term notes (MTNs), which US corporations began to issue in the early 1970s. In recent years outstanding corporate MTNs have averaged about 14 per cent of corporate bonds. They tend to be issued by highly rated corporations and are distinctive in being issued through 'shelf registrations' rather than having a formal offering with the assistance of underwriters. In a shelf registration an issuer presents a menu of securities that it may choose to issue in a specified period, which allows it to have a closer correspondence between the time funds are needed and the time when securities are issued. MTNs range in maturity from nine months to 30 years.

A large off-shore 'Eurobond' market exists where governments and corporations issue bonds denominated in currencies that differ from the currency in the country where the security is issued. While recent data are unavailable, there was also a rapidly growing outstanding stock of EuroMTNs in the early 1990s. These large and expanding markets complicate the implementation of monetary policy in a country, because information about Euromarkets must be taken into account. International financial statistics often do not reveal the nationality of individuals issuing or holding securities in different countries.

The establishment of financial instrument futures markets in 1975 also modified the demand for bonds in investor portfolios. Short-term hedging and speculative positions are more inexpensively achieved in a futures market than they are by constructing forward cash flows through the assumption of long and/or short positions in a bond market.

A market for 'stripped' bonds, where all a bond's coupons are separated from the body of a bond and each coupon and the body (or principal) are traded as separate entities, emerged in 1982. The body of the bond and each coupon are traded as discount bonds. The market for stripped securities greatly expanded in February 1985 when the US Treasury adopted this private sector innovation by offering its own stripped securities in book entry form and was willing to reconstruct stripped securities beginning in May 1987. These innovations increased the attractiveness of Treasury securities and arguably lowered the cost of government borrowing. The innovation is important because discount bonds are especially convenient for matching expected cash flows from other assets and liabilities and thus hedging against fluctuations in interest rates. Because discount bonds make no interest payments they are sometimes called 'zeros' in the financial press.

During the 1980s, a new technique emerged that broke the linkage between the choice of fixed or floating interest rates paid by a bond issuer and the form in which interest is received by a bondholder. A simple (plain vanilla) bond 'swap' is a transaction in which the holder of a bond trades a fixed interest-rate stream for a floating interest-rate stream. Thus, a borrower can issue a fixed-rate bond to an investor who prefers floating-rate securities, because the latter can simultaneously execute a swap with a third party. Such transactions facilitate marketing of securities in imperfectly competitive markets. Swaps also allow investors to change the currency unit in which an interest stream is denominated from, for example, euros to US dollars. They can also be used to change the base of a floating interest rate bond from, say, the US Treasury bill rate to dollar-denominated Libor, the London interbank offer rate.

Swaps and put and call options are early forms of 'derivative' securities, which allow investors to create synthetic bonds that effectively increase the stock of conventional corporate bonds, as can be inferred from Stoll (1969). A derivative security's value is conditional on the price or price trajectory over time of another asset. In recent decades an enormous variety of 'structured' assets has been and continues to be created by combining derivatives and conventional assets such as bonds and MTNs. For a discussion see Zhang (1997).

Finally, automation in bond markets has reduced the costs of trading bonds and made them more convenient to hold. Most government bonds in the United States are no longer issued in certificate form; they are issued in book form and exist only as computer entries. They are readily transferable in a computer and can be lent or sold at low cost whenever a borrower requires cash. By making bonds more reversible, automation has reduced the distinction between bonds and outside money, a distinction that is crucial for the success of central-bank open market operations.

See Also

- ▶ Fiat Money
- ▶ Miller, Merton (1923–2000)
- ▶ Modigliani, Franco (1918–2003)
- ▶ Monetary Economics, History of
- ▶ Residential Real Estate and Finance
- ▶ Sovereign Debt
- ▶ Third World Debt
- ▶ Tobin, James (1918–2002)

Bibliography

Becketti, S. 1988. The role of stripped securities in portfolio management. *Federal Reserve Bank of Kansas City Economic Review* 73: 20–31.
Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
Bulow, J., and K. Rogoff. 1988. The buyback boondoggle. *Brookings Papers on Economic Activity* 2: 675–704.
Bulow, J., and J. Shoven. 1978. The bankruptcy decision. *Bell Journal of Economics* 9: 437–456.
Cox, J., J. Ingersoll Jr., and S. Ross. 1981. The relation between forward prices and futures prices. *Journal of Financial Economics* 9: 321–346.
Crabbe, L. 1993. Anatomy of the medium-term note market. *Federal Reserve Bulletin* 79: 751–768.
Modigliani, F., and M. Miller. 1963. Corporate income taxes and the cost of capital: A correction. *American Economic Review* 53: 433–443.
Robinson, J. 1951. The rate of interest. *Econometrica* 19: 92–111.
Smith, W. 1970. The role of government-sponsored intermediaries. In *Housing and monetary policy*, Conference series no. 4, 86–101. Boston: Federal Reserve Bank of Boston.
Stoll, H. 1969. The relationship between put and call option prices. *Journal of Finance* 24: 801–824.
Tobin, J. 1963. An essay on the principles of debt management. In *Fiscal and debt management policies*, Prepared for the Commission on Money and Credit, 143–218. Englewood Cliffs: Prentice-Hall.
Wrase, J. 1997. Inflation-indexed bonds: How do they work? *Federal Reserve Bank of Philadelphia Business Review* 3–16.
Zhang, P. 1997. *Exotic options: A guide to second generation options*. Singapore: World Scientific Publishing.

Books, Economics of

Frederick van der Ploeg, Marcel Canoy and Jan van Ours

Abstract

The tensions between books as expressions of culture and books as profitable products are analysed using insights from the theory of industrial organization. To stimulate the diversity of books on offer, maintain the density of bookshops and to promote reading, governments grant fixed price monopolies, subsidize authors, levy a lower consumption tax on books, and provide public libraries and education. Market structures and government policies vary widely and there is no case for harmonizing European book policies. The book market is innovative in solving its problems. The main task of the government is to promote reading.

Keywords

Books; Economics of; Business stealing; Cross-subsidy; Cultural policy; Experience goods; Market failure; Monopolistic competition; Non-price competition; Payola; Product differentiation; Product life cycle; Retail price maintenance

JEL Classifications

L1; Z11

The market for books is characterized by the laws of demand and supply. However, the availability of a diverse supply of quality books is also an objective of cultural policy. This, combined with market failures, may provide grounds for government intervention as discussed for the arts in general in van der Ploeg (2006). Here we focus mainly on the market for general books, paying special attention to cultural books, leaving aside educational and scientific books. Governments influence book markets through subsidies for libraries, authors and publishers, tax concessions on the sale of books, and laws concerning the pricing of books. Apart from stimulating reading, it is not clear what role there is for government intervention. After all, the book market invents solutions to specific problems (contracts for authors, literary agents, gatekeeping by publishers, joint distribution by wholesalers cooperating on distribution, agreements

The authors would like to acknowledge that much of this article is based on Canoy et al. (2006), which also contains more details on the stylized facts and references

concerning stocks between retailers and publishers, joint publicity, best-seller lists, reviews, and so on). The book market flourishes in production of book titles, but not in reading.

Stylized Facts

About half of Portuguese adults never read a book. This is in sharp contrast with the 20 per cent of readers in Belgium, Denmark, Italy and Norway who similarly do not read books. Reading is popular in Finland, Sweden and Switzerland where about 90 per cent of adults read. Nevertheless, even in Sweden almost 30 per cent failed to read a book during 2003. Although in most countries a majority of adults read, there are large numbers of people who never read a book.

At the low end of the distribution of book titles across countries is the United States, with 24 titles per 100,000 inhabitants, and only six of which concern arts and culture. At the high end, Denmark produces 275 titles per 100,000 inhabitants, of which 80 are devoted to the arts and culture. Most titles per inhabitant are produced in Scandinavia, in Switzerland, and in the United Kingdom. Relatively few titles are produced in Italy, Japan, Greece and Australia.

The typical average annual number of books sold per inhabitant is about five to six. The exceptions at the lower end are Portugal and Sweden with 2.6 and 3.6 books per inhabitant, while at the high end France has 6.9. Publishers' revenues from sales vary from 20 euros per inhabitant in Greece to 115 euros in Finland. In most countries the revenue from book selling is 40–60 euros per inhabitant. The largest industries are located in the United States, Germany, the United Kingdom, France and Italy. In 2001, total value added of the book publishing industry was about 0.11 per cent of GDP with some 140,000 employees in the EU-15. The industry is stable in terms of turnover and per capita sales.

The number of books available through public libraries is low in Greece, Italy, Portugal and Spain, but much larger in Denmark, Finland and Sweden. The number of loans per inhabitant correlates nicely with the number of books available. It ranges from less than one in Greece, Portugal, Spain and Switzerland to at least ten in Denmark, Finland and the Netherlands. Differences in book-reading frequency are large. Reading a book daily varies from about a quarter of all adult males in Australia, Canada, Ireland, Sweden, Switzerland, the United Kingdom and the United States to a mere five per cent for Portuguese male adults. In most countries 10–20 per cent of adult males read daily. Females read much more than males, less so in Belgium (Flanders) and Portugal and more so in Australia, Canada, Denmark and the Netherlands.

The level of education is an important determinant of reading habits but no systematic cross-country evidence is available. However, in France 62, 78 and 92 per cent of lower-, medium-, and higher-educated individuals, respectively, read at last one book during the year 2003. There is not much cross-country information concerning trends in reading. However, in the Netherlands there is a clear downward trend in book-reading. Furthermore, fewer people indicate that they read books, though the average time spent reading has hardly changed. All readers irrespective of gender or country spend on average 6.5–8 hours per week reading books. In Europe, people spend most of their leisure time watching television. In the United States trends suggest that Internet use is increasing, mainly at the expense of watching television rather than reading.

Finland, Denmark, Ireland and Switzerland produce more than 200, the United Kingdom almost 200, Spain about 150 and the United States circa 25 book titles per 100,000 inhabitants per year. Although the number of titles produced has increased steadily in most countries, the number of publishers is stable. The average size of a publishing enterprise in the EU is small. Most enterprises publish only between 20 and 40 titles per year. The percentage of books published on arts and literature vary from 20 to 50 per cent across countries. Differences in the number of titles published may be related to economic prosperity, to the educational level of the population, or to population density. The empirical evidence for this is mixed. For example, a rich country like the United States publishes fewer titles per capita than some poorer southern European countries.

Total European book sales amounted to 27 billion euros in 2000. The biggest market is Germany with some 9.5 billion euros. Both Germany and the United Kingdom are strong exporters of books to countries that share their languages. Other large book markets are found in France, Spain and Italy. During the first two years of the 21st century, the United Kingdom book publishing industry has grown to be the largest in Europe. In contrast, there has been a decline in Germany. About half the revenues of publishers in most countries come from general books. Most sales are through retail channels (trade), except in the United States. In some countries there are strong retailers, but in others there are many independent bookshops. In France, the multimedia retailer Fnac accounts for around 15 per cent of sales. In Italy Feltrinelli commands 25 per cent of the retail market. However, in Germany, the largest bookseller, Thalia, has only three per cent of the market and there are many small independent bookshops. The largest retailers in the United Kingdom in 1998 were Waterstones and W.H. Smith with 20 per cent and 18 per cent of the market respectively. The United States book industry has limited opportunities for growth in a mature market, and competition is focused on growth through market shares. The United States has seen consolidation among retail chains. Barnes and Noble command 30 per cent of the market and independent booksellers struggle.

The share of book clubs is high in Australia (26 per cent), about 15–20 per cent in Denmark, Finland, France, and Sweden and low in Italy, the United Kingdom and the United States. Although Internet sales have grown in importance, they are still small. In the United Kingdom around 17 per cent of book sales go through Internet retailers, a percentage that is no longer thought to be growing very fast. For Germany estimates suggest between four and five per cent of sales are made through Internet retailers, although recent growth has been much faster than in the United Kingdom. Some reports have estimated Internet sales in France and Italy at 1–1.5 per cent. Spain has even lower Internet sales than France. Internet is mainly used as a channel for books and so far not for digital products. For example, E-books are not sold much in the European market. In the United States E-books are more important; over 7,000 titles were published in 2003 while over 1.3 million E-books were sold. Concentration of firms in the worldwide online book market is high, with 60 per cent for Amazon.com.

The Book Market Functions Well

According to Caves (2000) cultural goods are characterized by *nobody knows* (uncertain demand), *time flies* (short period of profitability), *infinite variety* (horizontal differentiation) and *A-list and B-list* (vertical differentiation). Beck (2003) adds spontaneous purchases of books, non-convexities in production with large fixed costs and small marginal costs, and free entry for the book trade. A book is a private good, since its consumption is rival and excludable. This suggests there is no fundamental market failure. Books can be borrowed by other people. However, if this yields utility to the owners, there is no market failure. The market for books has a traditional supply chain: production, wholesale, distribution and retail. In each part of the chain there is competition between private entrepreneurs. Government provision occurs only with libraries, but that does not exclude competition between private firms in the rest of the chain. There is substantial product differentiation in each part of the chain, which generates niche markets. Branding is important. Making a new product successful often requires substantial investment and innovation. This includes accepting that some products will never make it.

Most parts of the supply chain have a fairly large number of players. Consumers of books can easily switch from one product to the other. The book market knows relatively few consumer lock-ins, which helps the market to function properly. Transparency adds to that effect. Even though books are experience goods, author reputation, book reviews, book clubs and word-of-mouth ensure transparency. The book market is also dynamic: there is innovation, market shares fluctuate and there is entry and exit. All this suggests that the book market should not be exempted from competition law.

Books Occupy Niches, More So than Publishers

The book market is characterized by monopolistic competition along the lines of Dixit and Stiglitz (1997), since (*a*) products are differentiated; (*b*) firms set the price of the goods; (*c*) the number of sellers is large and each firm disregards the effects of its price decisions on the actions of its competitors; (*d*) entry is unrestricted. There thus exists a trade-off between efficiency (exploiting scale economies by producing more of the same product type) and diversity. Consumers love variety, but variety comes at a cost and the market becomes less transparent. Since firms do not take the potential downside of the variety decisions of other firms into account (the business stealing effect), there could be a market failure and optimal product diversity is not guaranteed. But in the book market consumers do not engage in repeated purchases in the same way as they do for, say, cereals. This greatly reduces possibilities for exploiting economies of scale, especially in the light of *nobody knows.* This does not mean that the book market can never have too much variety, but the argument then rests on lack of transparency and not on economies of scale. The book market does not have repeated entry by publishers with each publisher filling a niche. It is books that occupy niches, not publishers. Publishers have a portfolio of authors and books that serve as a way of risk-smoothing. Some books make it while others do not, but publishers have difficulties either of forecasting the success or are happy to accept differences in success out of cultural motives. Additional complexities arise for two other reasons. First, the book market is characterized by the fact that a single product (a book) has a very short life cycle. This leads to high initial prices followed by discounts. Second, publishers face a trade-off between risk-smoothing and specialization. A science fiction publisher has a competitive edge over nonspecialized publishers, but faces the risk that its clients might switch to video games.

A publisher thus has a quickly changing portfolio of books. Its strategy consists of deciding on the portfolio (trading off risks and specialization) and on the prices of the portfolio. Multi-product firms in a monopolistic competitive market face the decision whether to engage in new product lines (exploiting economies of scope) or not (reducing cannibalization). This is akin to the decision by a publisher whether to employ a new author in the same field as his current portfolio. This trade-off, combined with variety in publishers' 'love for culture', leads to a mix of publisher types. There are specialized publishers, small publishers and large publishers. This has been the case for many years in many countries.

The Book Market Plays Into Special Features of Books

Books have some special features. First, books are experience goods as one only appreciates the value after reading the book. Second, books are characterized by high fixed and low marginal costs. Third, some books are extremely successful, while most are unsuccessful. Success is hard to forecast and sometimes leads to 'winner takes all' economics as developed by Rosen (1981). Booksellers and publishers thus cross-subsidize higher-risk books with profits on other books. These potentially welfare-enhancing cross-subsidies can be thwarted by non-branch shops (for example, supermarkets) which sell only the bestsellers. Fourth, the opportunity costs of reading a book (that is, time) typically outweigh the price of a book. This contributes to a low price elasticity compared with other goods. The evidence suggests that the market for books other than best-sellers is price-inelastic, probably because most readers have high incomes or buy books for study purposes. Fifth, reading a book can be viewed as a private investment in culture rather than consumption. Sixth, there is an (almost) free substitute for buying books, namely, libraries. However, the quality of the service in bookshops and libraries is not the same, which makes substitutability imperfect. Seventh, books have cultural value. Books may also have option, existence and bequest value, and contribute to national identity, social cohesion, national prestige and the development of criticism and

experiments. None of these values is (fully) reflected in the price, so the total value of books is higher than what has been paid.

Still, the market need not fail, since publishers, booksellers and authors find solutions to cope with these special features. The book market is relatively simple compared with other cultural markets (Caves 2000). First, there is the *motley crew* property. A play or movie involves a complex set of different professionals to interact. The success of the play or movie crucially depends on how these different professionals get along. Many parts of the chain have the possibility to break it and kill the project. This leads to a complex set of contracts and other institutions, largely unnecessary and therefore absent in the book industry. Second, the *nobody knows* and *time flies* principles are even more applicable to a play or movie than to a book. Third, the production costs of a play or movie are much higher than those of a book.

Authors and publishers share the risk associated with the *nobody knows* and *time flies* principles. Authors get a percentage of the sales (typically ten per cent) and a split of the gross profits (typically 58–42) between author and publisher. Only celebrity authors receive bigger advances. While celebrity authors do reduce the risk of publishers somewhat, there are also serious large-scale flops. Some 70 per cent of former US President Bill Clinton's *Between Hope and History* were returned from bookstores as unsold (Caves 2000).

Changing the terms of the contract either in favour of the author or the publisher can lead to misallocations. A higher fee for the publisher leads to a higher number of published books, since it becomes more lucrative to publish books and there still exists a reservoir of authors wanting to accept lower fees (Caves 2000, p. 57). However, there will be less commercial success per book on average and lower quality as good authors may spend their time on more profitable activities. This could be justified if the perception is that there is a lack of supply of books. There is no evidence of that, however (the contrary is more likely). A higher percentage for the authors implies higher risk for the publisher, fewer books and fewer possibilities for new authors.

Incentives differ between publishers and authors. Publishers want to maximize profits, while many authors want to maximize sales and impact as they can often supplement their royalties with other income from lectures, TV, film, and so on. With globalization and the Internet some authors obtain superstar incomes by using the media to leverage their incomes. Many new authors find their way into the book market. In addition, sales of a novel increase the probability of future sales, a factor that influences an author more than the publisher. Authors may thus want to use agents. There is no marketplace for the literary reputations of new authors. The chance that a publisher accepts a manuscript is extremely low; Caves (2000) mentions one in 15,000 for novels. Agents reduce the cost of publishers by filtering out good and bad manuscripts. The publisher can then use the reputation of a good agent as a proxy for quality.

Nobody knows and *time flies* create problems with stocks in retail outlets. If a book does not perform, the retailer wants to dump stocks as shelf space is scarce and new potentially successful books are looming. Market solutions to this problem include second-hand sales shops, sales of remainders, pricing strategies and policies that aim at sharing risks between publishers and retailers. Book retailers also have a right to return books for full credit. They can further reduce risks by smart wholesaling agreements. There are distinct differences in market shares of wholesale firms in Europe. In France, Finland, Denmark and the Netherlands the wholesale market is concentrated, but in Anglo-Saxon countries wholesale is less concentrated. If publishers are larger, it is worthwhile for them to vertically integrate into distribution. In sum, the market seems fairly able to solve the coordination problems needed to sort out the economies of scale.

There also exists a trade-off between exploiting economies of scale in retail and other policy goals. Examples are the reduction of transport costs for consumers or equity 'universal service' type of arguments. Various trends such as the Internet tilt towards scale. Books are easy to transport and personal contact with the seller is not always needed. In fact, interactive service and

personal advice from Internet bookstores is often excellent.

Books are experience goods, so consumers have difficulty in deciding which book to buy. Book reviews in newspapers and the Internet, best-seller lists, book clubs, prizes and awards, and word of mouth facilitate choices. The market for information does not seem to fail except perhaps for payola (Caves 2000). Payola is a system where the author (or his agent) 'bribes' a gatekeeper to influence his choices (as with pop music on radio). For example, an author may buy many copies of his own book in order to be high on the best-sellers lists, or chain bookstores may offer deals to book publishers to selectively display books in eye-catching positions. The problem is that payola threatens the objectivity of gatekeepers.

Does the book market achieve cultural goals such as (i) a diverse supply of cultural book titles and genres; (ii) access of books for all in term of price and distance by having sufficient density and variety of (high-quality) retailers? Since books are rival and excludable, the book market should require less government interference. With the Internet one may expect a demand-driven growth in the sale of selected parts of handbooks and guidebooks. Because books are reproductive cultural goods, large-scale distribution of books is easier than for non-reproductive forms of art. The market thus produces a large variety of books, with prices that are low enough (with libraries as a fallback as well) to make books available to everybody interested. If retailers are unsuccessful in dealing with stock risks, there may be too few cultural books, too little reading or too many authors.

Should the Government Tolerate Retail Price Maintenance?

One reason to intervene is to protect a dense network of well-stocked, high-quality bookshops and stimulate the publication of a large variety of books. Indeed, the number of high-quality bookshops is decreasing in many countries. This happens if it does not pay to invest too much in variety in low-selling books. Monopoly profits and cross-subsidies from profitable to less profitable books may allow bookshops to store a greater variety of books and publishers to take more risks. The current practice in many European countries of a fixed book price (FBP) in combination with a variety of subsidies handed out by literary funds is often motivated by these considerations. Critics argue that a FBP or subsidies for high-brow books may harm reading on the part of the general public, since monopoly prices and cross-subsidies for less popular books are paid for by ordinary people reading popular books. Furthermore, subsidies for authors, translators, bookshops and publishers are paid for by ordinary people who may not be interested in more culturally valuable books or high-quality bookshops.

When considering policy instruments for reaching cultural objectives, there are at least two trade-offs. The first is between efficiency and density and distance. Increasing the scale of booksellers can enhance efficiency, but leads to longer travelling time for consumers. The second trade-off is between efficiency and cultural goals. Diversity of books in a bookstore may conflict with productive efficiency. The optimal choice of policy instruments depends on culture-political preferences and on country-specific characteristics that determine the market outcome. For example, a large 'language size' generates market outcomes where cultural objectives are more easily achieved. This is why the United States, Australia and Canada do not have policies aimed at the book market. Harmonizing book policies in Europe is not necessarily a good idea. Governments may wish to stimulate reading of worthwhile books, production of a diverse menu of titles and/or an extensive network of high-quality bookshops.

The FBP involves retail price maintenance by which the publisher reserves the right to set the retail prices of books. Since the publisher also influences wholesale prices, he effectively sets gross margins for retail outlets. The cultural merits ascribed to such agreements have reached almost mythical proportions in Europe. Since monopoly profits are higher than profits in competitive equilibrium, more titles are profitable and are

published or sold under the FBP than in competitive equilibrium. It is possible to print and sell extra books at low and almost non-increasing marginal cost, so the producer loss is likely to be small. Also, the price elasticity of the demand for books is small as a large part of the full cost of reading is the opportunity cost of time and thus monopoly profits are large. The FBP leads to more variety in book titles, but prices will be higher and sales of each title lower as discussed in van der Ploeg (2004). The welfare costs may in practice be much larger, since much of the profit is dissipated by unproductive rent-seeking along the lines of Tullock (1980).

The FBP also has dynamic costs. Price competition between retail outlets becomes impossible but it is also more difficult to vary prices in response to local conditions. A store on a remote island may want to charge more for the same book than a store in the capital, but under the FBP it cannot do so. Also, it is more difficult to vary prices for different types of customers or for different seasons. Some customers need no service and low prices, while others prefer service at a higher price. Most important is that the FBP discourages the development of innovative distribution channels, since realized cost savings cannot be passed on to customers. With the FBP, unconventional distribution channels (bookclubs, supermarkets, petrol stations, the Internet, and so on) have less of a chance. Against these costs there is the benefit that independent small bookshops may be able to recommend interesting books and order books from the publisher or distributor.

Potential Gains from Retail Price Maintenance

Even though the FBP eliminates price competition, non-price competition may intensify. For example, a bigger sale margin stimulates booksellers to give better service to customers (Holahan 1979; Mathewson and Winter 1998; Deneckere et al. 1997). With a bigger profit margin, it pays to spend more effort on service in order to get extra customers. If the extra service (more attractive presentation in bookshops, better information to customers, more promotion, and so on) generates more sales than the fallback in sales due to higher monopoly prices, the FBP may be desirable. Otherwise, the market fails to deliver sufficient service, because bookshops have an incentive to operate as free riders by offering discounts and expecting their customers to get their information and service elsewhere. Bookshops hardly refuse service or charge for information provided to people who in the end may not buy a book. Still, most customers rarely engage in such a strategy, as the costs of roaming around various bookshops seem high in relation to the possible discount one might obtain. Much of this service is already made available through publishers' advertisements or book reviews in newspapers and other media or on the Internet. In any case, it is questionable whether the demand for books really depends on service. Better service does not seem a good argument for supporting a FBP.

The book trade also argues that a bigger margin provides incentives for better- stocked bookshops. Booksellers may take over some of the inventory risks from publishers, so that more titles will be published. At the margin it is more profitable for retail outlets with relatively high costs to open up. This argument works only if customers want to purchase their books at particular high-cost bookshops. The gain in sales from these outlets may then offset the drop in sales resulting from higher monopoly prices. Although a dense network of bookshops may be desirable from a cultural point of view, this argument for the FBP is difficult to justify on grounds of market failure. Another popular argument is that higher margins encourage more retail outlets to put new book titles with uncertain sales prospects on their shelves. Given that there seems to be no problem for new authors to get their first book published, this is not a strong argument either. Marvel and McCafferty (1984) suggest that resale price maintenance may sustain a luxury image, but that seems more relevant for the markets for perfumes and jewellery than for books.

Is The Cross-Subsidy Argument Really Valid?

The novel *Endurance* by Ian McEwan is not a perfect substitute for *Il Nome della Rose* by Umberto Eco. They are different books, because the authors have different styles, the themes of the

two novels are different, and last but not least the original languages in which the books are written are different. Still, Umberto Eco's books are closer substitutes for the novels of Ian McEwan than, say, a cookbook or a travel book. On the other hand, Martin Amis may be a closer substitute than Umberto Eco for Ian McEwan. One must therefore leave the realms of homogenous goods and adopt a framework of Chamberlinian monopolistic competition in which books are imperfect substitutes. Publishers and booksellers carve out a niche and make monopoly profits, which enable them to recoup fixed costs. It is thus profitable to publish books. In fact, an important argument of the lobby of booksellers and publishers rests on imperfect competition. They argue that the FBP allows for cross-subsidies from best-sellers to less popular books and leads to a more diverse supply of book titles and bookshops. In addition, the book lobby suggests that publishing and stocking a large selection of books enhances reputation, yields economies of scope and satisfies the idiosyncratic taste of individual publishers and booksellers even though these arguments do not seem very strong (Canoy et al. 2006).

The cross-subsidy argument seems at first blush irrelevant. In competitive markets with imperfect information about the success of a product, it is common to invest in many products and reap a success on only a few. Even without a fixed horse price agreement, horse owners purchase lots of yearlings, many of which are subsequently sold to the riding school or the butcher if they do not win races. Similarly, in a market without FBP, publishers invest in new authors, just as horse owners invest in yearlings. Indeed, the industry's rule of thumb formulated by Denis Diderot in 1767 suggests that one out of ten new editions is a profitable success, four cover costs, and five make losses (Beck 2003). There are few barriers to new authors in the book market even though publishing is a risky business with only one- third of published books being profitable. The FBP then has all the welfare and political economy costs of a monopoly. This situation may arise if best-sellers are easily digestible, require little time to read and have high price elasticities of demand, while, say, poetry readings demand a lot of time and effort and have low price elasticities of demand. Indeed, anything worthwhile from a cultural point of view takes time and effort to appreciate and contributes to a low price elasticity of demand.

Non-fiction books (dictionaries, cookbooks, travel guides, textbooks, and so on) are likely to be close substitutes within each genre and will thus have high price elasticities. Fiction books (children books, mysteries, and so on) often have close substitutes (perhaps with the exception of *Harry Potter*), especially for the pocketbook versions of old titles, and thus high price elasticities. We do not expect large monopoly profits on such titles, and there is little room for cross-subsidies to books with a special or unique character. Such books have low price elasticities and generate high monopoly profits. If this is the situation, the cross-subsidy argument is likely to be wrong. The problem with a FBP is that there is no guarantee that publishers and booksellers will use the monopoly profits to make sure that more esoteric titles will be published and stocked in the stores. Monopoly profits may well be directed towards unproductive managerial slack.

Summing Up

In summary, a FBP may induce higher prices and fewer sales of any book title that is published. It may also hinder innovation and distribution, but more titles will be published and there will be more bookshops with a diverse assortment of titles. However, German data suggest that retail price maintenance does not facilitate above-average focal pricing where prices are bunched around focal points (Beck 2004). The lowering of production costs due to technological progress will benefit the diversity of books being published. In any case, many FBPs are of limited duration and characterized by sensible exceptions. The welfare costs are probably not very large, but may be reduced a little by reducing the term and coverage of the agreement. It may also be helpful to abolish certification and exclusive trade arrangements, scrap the fixed discount for recognized booksellers, and move to individual rather than vertical price agreements (see also Appelman and van den Broek 2002). Since educational and

scientific books typically have relatively low price elasticities and are more susceptible to monopoly abuse, it helps to exclude them from the FBP. As a dogma, the FBP diverts attention and energy away from making the book trade more innovative and customer-oriented. It may be more worthwhile to stimulate reading of a wide variety of books by investing in public libraries and education, subsidising authors to write books of high cultural value, translating the best books into other languages and promoting them abroad.

Other Public Policies

Stimulating Demand: Lower Value-Added Tax

The general consumption of books can be increased by lowering the specific value- added tax (VAT) rate on books. This is a general instrument, which is not well suited to direct at special books of literary value. The lower VAT on books applies to cookbooks as well as to poetry. This instrument is therefore mainly used to stimulate the purchasing and, it is hoped, reading of books. Administrative costs are low, since no apparatus of literary experts has to be called upon. All countries of Europe, except Denmark, use a reduced VAT rate as instrument to stimulate book purchases. The United Kingdom and Ireland even abolished VAT on books altogether. The European Commission misguidedly attempts to harmonize VAT rates on books, making it difficult for other member states to abolish VAT on books. The Commission fails to take account of the subsidiarity principle. Since the book trade, especially between the non-English speaking countries, hardly distorts the intra- European book trade, there is no danger of tax competition and no harm in countries pursuing their VAT policies on books independently of each other.

Stimulating Supply: Prizes and Grants for Writers and Subsidies for Bookshops

Governments and commercial sponsors do many things to encourage writers. There are many prestigious and less prestigious prizes for the best novelist, the best detective writer, the best poet, the best translator, and so on. All these are meant to encourage quality. More important, they might guide the uninitiated reader to better books. Book clubs, best-seller lists and book programmes on television also help in this respect. They also probably increase sales. Literary funds help struggling authors to make a living if their project is deemed to be of literary interest. Since only best-seller authors can make a living on royalties and related incomes, others may need some help, especially if their output has cultural value but is perhaps of less general interest. These policies are designed to stimulate quality rather than quantity. Sometimes subsidies for publishers of high-quality books may help as well (witness Sweden).

Many politicians attach cultural importance to a dense network of retail outlets. We have already noted that density seems to be falling in some countries, perhaps more in countries without a FBP; and concentration is increasing as well. From a cultural point of view this is bad news. Consumers have to travel further and there is less variety of bookshops. If the main objective of cultural policies is to increase the density of high-quality outlets, subsidies for high-quality bookshops may be more effective than the FBP. If they act as cultural centres in less populated areas, they may deserve public support.

Subsidizing in order to maintain well-stocked bookshops would probably prove an administrative nightmare, which may explain why there is not much experience of this. Subsidizing publishers to publish books of literary and cultural value would also seem to hinder the market mechanism and lead to adverse effects. In Sweden the government subsidizes in this manner roughly one-third of all fiction and one-fifth of books for children. However, Swedish retailers do not stock all titles since the government, rather surprisingly, does not require subsidized books to be offered for sale.

Concluding Remarks

The book market ensures reasonable cultural performance with little government intervention,

especially in large language areas. Yet there are differences between countries in reading, retail outlets, wholesale and production. Due to lack of data and research it is not easy to explain these differences. They may be due to differences in preferences, logistics, population density or public policies or to being stuck in the wrong equilibrium. One important trend is that people seem to read fewer books over time. Perhaps they are reading on the Internet or spending time on other cultural leisure activities. Here are some important areas for further research: investigating the relationship between production of titles, books sold and prices; using survey data to study the effects of personal characteristics of readers on market outcomes; analysing empirically differences between book and other cultural markets; and using industrial organization to understand pricing and stocking behaviour of publishers and retailers.

The book industry is characterized by relatively few market failures and these can be relatively easily corrected with market instruments. The book industry can fend well for itself, in contrast to opera, film or theatre, characterized by high production costs, high risk and complex interactions between a large number of different professionals. Even though there are obvious returns to scale, production costs are low. Thresholds for new authors, publishers and retailers are small, contracts are relatively simple and fairly uniform. The market is quite capable of inventing solutions for specific problems and public policies are not always called for, except perhaps to stimulate reading.

Nevertheless, there is a strong lobby for government intervention. Prizes and grants for authors, translators, publishers, bookshops, special VAT regimes for books, stimulating reading through public libraries, and the FBP are possible policy instruments. The standard case against the FBP is that book prices are higher and sales lower than under perfect competition. This hurts the interests of buyers, particularly those with lower incomes, since prices will be higher. One possible argument in favour is that the FBP may induce more and better-stocked bookshops and lead to publication of more marginal book titles. The cross-subsidy argument of the lobby in favour of the FBP is not convincing, however. First, even without the FBP, the market cross-subsidizes new authors and other risky projects in the hope of a possible best-seller. Second, even if this policy 'works', there is no accounting for what is done with the cross-subsidies and no democratic checks. Third, there is no guarantee that profits on best-sellers will be used to cross-subsidize less popular books. In fact, publishers and booksellers have an incentive not to do this. Fourth, if less popular books are less price elastic than popular books (perhaps as they take more time to read), monopoly profits on less popular books are higher and the cross-subsidy argument does not work. Fifth, even if cross-subsidization does occur, one should evaluate whether its cultural gains outweigh the distortionary costs of the FBP. Arguments put forward to defend the FBP, stressing improved service, better distribution and retail networks, and other forms of increased non-price competition, do not stand up to scrutiny either. The book industry produces many titles and new authors do not experience severe problems. The FBP may slow down or even stop the declining number of well-stocked bookshops outside big cities, but hinders sales through the Internet and supermarkets.

A comparison of policies towards the book industry in different European countries teaches us that harmonization is a bad idea. There is not much inter- European book trade, so that book policies hardly distort the single European market. Also, characteristics of book industry, cultural and social features and political preferences of the different countries of Europe differ substantially. It is therefore best to allow member states of the European Union to design their own book policies. For example, a FBP makes more sense for Greece than for the United Kingdom as it has a smaller 'language size' and fewer people have access to the Internet. Although there may be a problem of a 'race to the bottom' if VAT rates are not harmonized, tax competition seems pretty irrelevant for the book market. European countries should be free to lower or abolish VAT on books in order to promote reading.

Many of the privileges granted in the book industry will eventually be undermined by technical changes. Digital cameras, recording and editing equipment have made low budget radio and television as well as narrowcasting possible, thus undermining the monopoly power of public and other broadcasters. Similarly, the Internet has stimulated virtual book suppliers, printing and publishing on demand and E-books. Virtual dictionaries, encyclopedias and other handbooks have already overtaken, to a large extent, their physical counterparts. A dense network of well- stocked bookshops remains important. Some argue that the emergence of the Internet and the integration of books in smart product and digitized communication will lead to the disappearance of the printed book (Choi et al. 1998). While more retailing will take place through the Internet and new gadgets, for some people the physical bookshop, where one can feel the book and bump into surprise titles and people, will remain indispensable.

There are, however, trends that endanger books, the most important being that people read less and less. Some worry that the next generation will stop reading books altogether, but this may be too pessimistic. First, the population is aging so that more leisure time becomes available and the opportunity costs of reading decrease. Second, books are doing well. In 1947, some 85,000 books were in print in the United States, against 1.3 million in 1996. This is, in part, due to sharp reductions in production and printing costs. Third, there is no reason to believe that a cultural carrier as old as the book will suddenly disappear. Modern technology complements books rather than substitutes for them (Cowen 1998).

Each new development in the craft has led to outbursts of cultural pessimism that allegedly indicates the end of the book. Most of the developments have only improved the book business (Cowen 1998). Also, prices fell considerably and steadily. The future of the book market may look very different. E-books will replace parts of the market where E-reading already outperforms traditional reading. As for novels, nobody knows. Perhaps our children will read their novels directly from the screen.

See Also

- ▶ Art, Economics of
- ▶ Consumer Expenditure (New Developments and the State of Research)
- ▶ Internet, Economics of the
- ▶ Markets
- ▶ Product Differentiation

Bibliography

Allen, W., and P. Curwen. 1991. *Competition and choice in the publishing industry*. London: Institute of Economic Affairs.

Appelman, M.D., and A. van den Broek. 2002. *Boek en Markt. Effectiviteit en Efficientië van de Vaste Boekenprijs*. The Hague: SCP/CPB.

Appelman, M.D., and M.F.M. Canoy. 2002. Horses for courses: Why Europe should not harmonise its book policies. *De Economist* 150: 583–600.

Barro, R.J., and J.-W. Lee. 2000. *International data on educational attainment: Updates and implications*, CID Working Paper, vol. 42. Cambridge, MA: Harvard University.

Beck, J. 2003. *Monopoly vs oligopoly in the debate on retail price maintenance for books: A preliminary empirical result*. Berlin: Diplomarbeit/Humboldt University.

Beck, J. 2004. *Fixed, focal or fair? Book prices under optional retail price maintenance*, SP II 2004–15. Berlin: Wissenschaftszentrum.

Bertarelli, S., and R. Censolo. 2000. *Preference for novelty and price behaviour*. Working Paper, University of Bologna.

Bittlingmayer, G. 1992. The elasticity of the demand for books, resale price maintenance and the lerner index. *Journal of Institutional and Theoretical Economics* 148: 588–606.

Canoy, M., J. van Ours, and F. van der Ploeg. 2006. The economics of books. In *The handbook of the economics of art and culture*, ed. D. Ginsburgh and D. Throsby. Amsterdam: North-Holland.

Caves, R. 2000. *Creative industries, contracts between art and commerce*. Cambridge, MA: Harvard University Press.

Choi, S.-Y.C., D.O. Stahl, and A.B.. Whinston. 1998. Gutenberg and the digital revolution: Will printed books disappear? *Journal of Internet Banking and Commerce* 21: 21–21.

Clerides, S.K. 2002. Book value: Intertemporal pricing and quality discrimination in the US market for books. *International Journal of Industrial Organization* 20: 1385–1408.

Coser, L.A., C. Kadushin, and W.W. Powell. 1982. *Books: The culture and commerce of publishing*. Chicago: University of Chicago Press.

Cowen, T. 1998. *In praise of commercial culture*. Cambridge, MA: Harvard University Press.
Creemers, M. 1999. In de electronische boekhandel is informatie belangrijker dan prijs. *Holland Management Review* 66: 58–63.
Cummings Jr., M.C., and R.S. Katz. 1987. *The patron state*. New York: Oxford University Press.
de Grauwe, P., and G. Gielens. 1993. *De prijs van het boek en de leescultuur*, CES Working Paper, vol. 12. University of Leuven.
Deneckere, R., H.P. Marvel, and J. Peck. 1997. Demand uncertainty and price maintenance: Markdowns as destructive competition. *American Economic Review* 87: 619–641.
Dixit, A., and J.E. Stiglitz. 1997. Monopolistic competition and optimal product diversity. *American Economic Review* 67: 297–308.
Ecalle, F. 1988. Une évolution de la loi relative du 10 août au prix de livre. Economie et prévision 5/88*, Paris.*
Economisch Instituut voor het Midden- en Kleinbedrijf. 2001. *De Vaste Boekenprijs: Een Internationale 'Quick Scan*. Zoetermeer: EIM.
Epstein, J. 2001. *Book business: Publishing past, present and future*. New York: W.W. Norton.
European Commission. 2002. *Commission accepts undertaking in competition proceedings regarding German book price fixing*, Press release IP/02/461. Brussels: European Commission.
European Commission. 2004. *Book publishing. Publishing market watch. Sectoral report 2*. Brussels: European Commission.
Eurostat. 2001. *Eurostat yearbook 2002*. Luxemburg: European Commission.
Eurostat. 2002. *Statistics in focus, theme 4–24*. Luxemburg: European Commission.
Eurostat. 2004. *How Europeans spend their Time – Everyday life of women and men*. Luxemburg: Office for Official Publications of the European Communities.
Fishwick, F. 1989. Les implications économiques du Net Book Agreement. *Cahiers de l'Économie du Livre* 2: 4–31.
Fishwick, F., and S. Fitzsimons. 1998. *Report into the effects of the abandonment of the net book agreement*. Cranfield: Cranfield School of Management.
Foster, J.E., and A.W. Horowitz. 1996. Complimentarily yours: Free examination copies and textbook prices. *International Journal of Industrial Organization* 14: 85–99.
Goolsbee, A., and J. Chevalier. 2002. *Measuring prices and competition online: Amazon and Barnes and Noble*, Working Paper, vol. 9085. Cambridge, MA: NBER.
Greco, A.N. 1999. The impact of horizontal mergers and acquisitions on corporate concentration in the U.S. book publishing industry: 1989–1994. *Journal of Media Economics* 12: 165–180.
Greco, A.N. 2000. Market concentration levels in the U.S. consumer book industry: 1995–1996. *Journal of Cultural Economics* 24: 321–336.
Hjorth-Andersen, Chr. 2000. A model of the Danish book market. *Journal of Cultural Economics* 24: 27–43.
Holahan, W.L. 1979. A theoretical analysis of resale price maintenance. *Journal of Economic Theory* 21: 411–420.
Klein, D.C. 2000. Web strategies for professional publishers: Developing an information services portal. *Learned Publishing* 13: 83–94.
Landes, W.M., and R.A. Posner. 1989. An economic analysis of copyright law. *Journal of Legal Studies* 18: 325–363.
Latcovich, S., and H. Smith. 2001. Pricing, sunk costs, and market structure online: Evidence from book retailing. *Oxford Review of Economic Policy* 17: 217–234.
Marvel, H.P., and S. McCafferty. 1984. Resale price maintenance and quality certification. *RAND Journal of Economics* 15: 346–359.
Mathewson, F., and R. Winter. 1998. The law and economics of resale price maintenance. *Review of Industrial Organisation* 13: 57–84.
O'Hagan, J.W. 1998. *The state and the arts: An analysis of key economic policy issues in Europe and the United States*. North Hampton: Edward Elgar.
Ornstein, S.I. 1985. Retail price maintenance and cartels. *Antitrust Bulletin* 30: 401–432.
Ottavanio, G.I.P., and J.F. Thisse. 1999. *Monopolistic competition, multiproduct firms and optimum product diversity*, Discussion Paper, vol. 9919. Louvain: CORE.
Plant, A. 1934. The economic aspects of copyright in books. *Economica* 1: 167–195.
Ringstad, V. 2004. On the cultural blessings of fixed book prices: Fact or fiction. *International Journal of Cultural Policy* 10: 351–365.
Rosen, S. 1981. The economics of superstars. *American Economic Review* 71: 848–858.
Rürup, B., R. Klopfleisch, and H. Stumpp. 1997. *Ökomomische Analyse der Buchpreisbindung*. Frankfurt: Hessischer Verleger- und Buchhändler-Verband/ Büchhändler-Vereinigung.
Sutton, J. 2000. *Marshall's tendencies. What can economists know?* Cambridge, MA: Leuven University Press/MIT Press.
Szenberg, M., and E. Youngkoo Lee. 1994. The structure of the American book publishing industry. *Journal of Cultural Economics* 18: 313–322.
Throsby, D.C. 2001. *Economics and culture*. Cambridge: Cambridge University Press.
Tietzel, M. 1995. *Literaturökonomik*. Tübingen: Mohr.
Tirole, J. 1998. *The theory of industrial organization*. Cambridge, MA: MIT Press.
Tullock, G. 1980. Efficient rent seeking. In *Towards a theory of the rent-seeking society*, ed. J.M. Buchanan, R.D. Tollison, G. Tullock, and G. Tullock. College Station: Texas A&M Press.
Uitermark, P.J. 1986. Verticale prijsbinding van boeken; concurrentie en cultuur: een aanzet tot analyse. In *De*

Vaste Boekenprijs: Pro's en Contra's. Leiden: Instituut voor Onderzoek van Overheidsuitgaven.
van der Ploeg, F. 2004. Beyond the dogma of the fixed book price agreement. *Journal of Cultural Economics* 28: 1–20.
van der Ploeg, F. 2006. The making of cultural policy: A European perspective. In *The handbook of the economics of art and culture*, ed. D. Ginsburgh and D. Throsby. Amsterdam: North-Holland.
van Ours, J.C. 1990. De Nederlandse boekenmarkt tussen stabiliteit en verandering. *Massacommunicatie* 18: 22–35.
Whyte, J.L. 1994. Breaking the bookshop cartel. *Long Range Planning* 27: 75–87.
Yetkiner, I.H., and C. Horvth. 2000. *Macroeconomic implications of virtual shopping: A theoretical approach*. Groningen: University of Groningen Press.

Bootstrap

Joel Horowitz

Abstract

The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data. It is often much more accurate in finite samples than ordinary asymptotic approximations are. This is important in applied research, because the familiar asymptotic normal and chi-square approximations can be very inaccurate. When this happens, the difference between the true and nominal coverage probability of a confidence interval or rejection probability of a test can be very large, and inference can be highly misleading. The bootstrap often greatly reduces errors in coverage and rejection probabilities, thereby making reliable inference possible.

Keywords

Asymptotic distribution; Asymptotic refinements; Bias reduction; Bootstrap; Conditional Kolmogorov test statistic; Edgeworth approximations; Maximum score estimator; Monte Carlo simulation; Probability; Probit models; Statistical inference; Statistics and economics; Subsampling; Tobit model

JEL Classifications

C15

The bootstrap is a method for estimating the sampling distribution of an estimator or test statistic by resampling one's data. It amounts to treating the data as if they were the population for the purpose of evaluating the distribution of interest. Under mild regularity conditions, the bootstrap yields an approximation to the distribution of an estimator or test statistic that is at least as accurate as the approximation obtained from 'ordinary' or first-order asymptotic theory. Thus, the bootstrap provides a way to substitute computation for mathematical analysis if calculating the asymptotic distribution of an estimator or statistic is difficult. Moreover, the bootstrap is often more accurate in finite samples than first-order asymptotic approximations are but does not entail the algebraic complexity of higher-order expansions. Thus, it can provide a practical method for improving upon first-order approximations. Such improvements are called 'asymptotic refinements'.

The bootstrap is of considerable importance in applied research. Many important statistics in econometrics have complicated asymptotic distributions that depend on nuisance parameters and, therefore, cannot be tabulated. Examples include the conditional Kolmogorov test statistic of Andrews (1997) and Manski's (1975, 1985) maximum score estimator for a binary-response model. The bootstrap and related resampling techniques provide practical methods for estimating the distributions of such statistics. In other cases, the statistic of interest has a familiar distribution but with a complicated standard error that is difficult to work with analytically (for example, Horowitz and Manski 2000). Again, the bootstrap provides a practical method for carrying out inference.

The bootstrap's ability to provide asymptotic refinements is especially important in applied research. First-order asymptotic approximations (for example, asymptotic normal and chi-square approximations) can be very inaccurate with the sample sizes that are found in applications. When

this happens, the difference between the true and nominal coverage probability of a confidence interval (error in the coverage probability or ECP) can be very large. Similarly, the difference between the true and nominal probability that a test rejects a correct null hypothesis (error in the rejection probability or ERP) can be very large. Consequently, inference based on first-order asymptotic approximations can be highly misleading. White's (1982) information matrix test is a well-known example of this. There are many others. The bootstrap often greatly reduces the ECPs of confidence intervals and ERPs of tests, thereby making reliable inference possible.

Bias reduction is another use of the bootstrap's ability to provide asymptotic refinements. It is not unusual for an asymptotically unbiased estimator to have a large finite-sample bias. This may cause the estimator's finite-sample mean-square error to be very large. The bootstrap can be used to reduce the estimator's finite-sample bias and, thereby, its finite-sample mean-square error.

The bootstrap has been the object of research in statistics since its introduction by Efron (1979). The results of this research are synthesized in the books by Beran and Ducharme (1991), Davison and Hinkley (1997), Efron and Tibshirani (1993), Hall (1992), Mammen (1992), and Shao and Tu (1995). Hall (1994), Horowitz (1997, 2003), Maddala and Jeong (1993), and Vinod (1993) provide reviews with an econometric orientation. Horowitz (2001) provides a detailed discussion of the theory and use of the bootstrap in econometrics.

This article assumes that the data are an independent random sample from some distribution. Horowitz (2001) and Lahiri (2003) discuss bootstrap methods for time-series data.

How the Bootstrap Works

This section explains why the bootstrap works and how it is implemented in simple settings. The estimation problem to be solved may be stated as follows. Let the data, $\{X_i : i = 1, \ldots, n\}$, be a random sample of size n from a probability distribution whose cumulative distribution function (CDF) is F. Let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic (that is, a function of the data), possibly a test statistic. Let $G_n(\tau, F) = P(T_n \leq \tau)$ denote the exact, finite-sample CDF of T_n. Usually, $G_n(\tau, F)$ is a different function of τ for different distributions F. An exception occurs if $G_n(\tau, F)$ does not depend on F, in which case T_n is said to be *pivotal*, but pivotal statistics are not available in most applications. Therefore, $G_n(\tau, F)$ cannot be calculated if, as is usually the case in applications, F is unknown. The bootstrap is a method for estimating $G_n(\tau, F)$ or features of it such as its quantiles when F is unknown.

First-order asymptotic distribution theory is another method for estimating $G_n(\tau, F)$. The asymptotic distributions of many econometric statistics are standard normal or chi-square, possibly after centring and normalization, regardless of the distribution from which the data were sampled. Such statistics are called *asymptotically pivotal*, meaning that their asymptotic distributions do not depend on unknown population parameters. Let $G_\infty(\tau, F)$ denote the asymptotic distribution of T_n. If T_n is asymptotically pivotal, then $G_\infty(\cdot, F) = G_\infty(\cdot)$ does not depend on F. Therefore, if n is sufficiently large, $G_n(\cdot, F)$ can be estimated by $G_\infty(\cdot)$ without knowing F. This method for estimating $G_n(\cdot, F)$ is often easy to implement and is widely used. However, $G_\infty(\cdot)$ can be a poor approximation to $G_n(\cdot, F)$ with samples of the sizes encountered in applications.

The bootstrap provides an alternative approximation to $G_n(\cdot, F)$. Whereas first-order asymptotic approximations replace the unknown distribution function G_n with the known function G_∞, the bootstrap replaces the unknown distribution function F with a consistent estimator such as the empirical distribution function of the data. Let F_n denote the estimator of F. The bootstrap estimator of $G_n(\cdot, F)$ is $G_n(\cdot, F_n)$. Usually, $G_n(\cdot, F_n)$ cannot be evaluated analytically. It can, however, be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which random samples are drawn from the data. Thus, the bootstrap is usually implemented by Monte Carlo simulation. The Monte Carlo procedure for estimating $G_n(\tau, F_n)$ is:

Step 1: Generate a bootstrap sample, $\{X_i^* : i = 1,...,n\}$ by sampling the estimation data randomly with replacement.

Step 2: Compute $T_n^* = T_n(X_1^*,...,X_n^*)$.

Step 3: Use the results of many repetitions of steps 1 and 2 to compute the empirical probability of the even $T_n^* \leq \tau$ (that is, the proportion of repetitions in which this event occurs).

If T_n is a test statistic, then the bootstrap can be used to estimate its critical value. Consider a test that rejects the null hypothesis, H_0, if $|T_n|$ is too large. The exact α-level critical value, $z_{n,\alpha/2}$, is the solution to $G_n(z_{n,\alpha/2}, F) - G_n(-z_{n,\alpha/2}, F) = 1 - \alpha$. Unless T_n is pivotal, however, this equation cannot be solved in an application because F is unknown. Therefore, the exact, finite-sample critical value cannot be obtained in an application if T_n is not pivotal. The bootstrap replaces F with F_n. Thus, the bootstrap critical value, $Z_{n,\alpha/2}^*$, solves $G_n\left(Z_{n,\alpha/2}^*, F_n\right) - G_n\left(-z_{n,\alpha/2}^*, F_n\right) = 1 - \alpha$. This equation usually cannot be solved analytically, but $Z_{n,\alpha/2}^*$ can be estimated with any desired accuracy by Monte Carlo simulation. To illustrate, suppose, as often happens in applications, that T_n is an asymptotically standard normal, Studentized estimator of a parameter θ whose value under H_0 is θ_0. That is, $T_n = n^{1/2}(\theta_n - \theta_0)/s_n$, where θ_n is the estimator of θ, $n^{1/2}(\theta_n - \theta_0) \rightarrow^{d} N(0, \sigma^2)$ under H_0, and s_n^2 is a consistent estimator of σ^2. Then the Monte Carlo procedure for computing $Z_{n,\alpha/2}^*$ is:

Step 1. Use the estimation data to compute θ_n.

Step 2. Generate a bootstrap sample of size n by sampling the data randomly with replacement. Compute the estimators of θ and σ from the bootstrap sample. Call the results θ_n^* and s_n^*. The bootstrap version of T_n is $T_n^* = n^{1/2}(\theta_n^* - \theta_n)/s_n^*$.

Step 3. Use the results of many repetitions of step 2 to compute the empirical distribution of $|T_n^*|$. Set $Z_{n,\alpha/2}^*$ equal to the $1 - \alpha$ quantile of this distribution.

A test based on $|T_n|$ and the bootstrap critical value rejects H_0 at the α-level if $|T_n| > Z_{n,\alpha/2}^*$. A symmetrical 1 - α confidence interval for θ based on the bootstrap critical value is $\theta_n - z_{n,\alpha/2}^* s_n \leq \theta \leq \theta_n + z_{n,\alpha/2}^* s_n$. For reasons that are explained in Section 2, use of the bootstrap critical value $z_{n,\alpha/2}^*$ instead of the critical value based on the asymptotic normal distribution can greatly reduce the ERP of a test of a hypothesis about θ and the ECP of a confidence interval for θ.

Since F_n and F are different functions, $G_n(\cdot, F_n)$ and $G_n(\cdot, F)$ are also different functions unless T_n is pivotal. Therefore, the bootstrap estimators $G_n(\cdot, F_n)$ and $z_{n,\alpha/2}^*$ are only approximations to the exact finite-sample CDF and critical value of T_n, $G_n(\cdot, F)$ and $z_{n,\alpha/2}$. However, F_n is close to F when n is large. Therefore, if G_n is a sufficiently smooth function, $G_n(\cdot, F_n)$ will be close to $G_n(\cdot, F)$. Moreover, we can expect $z_{n,\alpha/2}^*$ to approach $z_{n,\alpha/2}$ as $n \rightarrow \infty$. In other words, the bootstrap provides an approximation to the sampling distribution and critical value of T_n that becomes increasingly accurate as n increases. This property of the bootstrap is called *consistency*. Beran and Ducharme (1991) and Mammen (1992) give formal conditions under which the bootstrap is consistent. Horowitz (2001) gives some econometrically relevant examples in which the bootstrap is not consistent and, therefore, cannot be used to estimate the distribution of a statistic. These include Manski's maximum score estimator, the distribution of a parameter on the boundary of the parameter set, and estimation of the maximum of a sample.

When the bootstrap is inconsistent (that is, $G_n(\cdot, F_n) - G_n(\cdot, F)$ does not converge to 0), subsampling procedures can be used to estimate $G_n(\cdot, F)$. One approach to subsampling consists of drawing samples of size $m < n$ by sampling the data randomly *without* replacement. This produces random samples from the true population distribution of the data, F, not the empirical distribution, F_n, from which bootstrap samples are drawn. Consequently, subsampling yields a consistent estimator of $G_n(\cdot, F)$, even when the bootstrap does not. Politis et al. (1999) describe the theory of subsampling and methods for implementation. Subsampling is consistent in all known settings of practical importance, so it is much more widely

applicable than the bootstrap. The price of this versatility, however, is reduced accuracy. The approximation provided by subsampling is typically less accurate than that provided by first-order asymptotic distribution theory, and subsampling can be much less accurate than the bootstrap when the bootstrap is consistent.

Asymptotic Refinements

The bootstrap provides asymptotic refinements for statistics that are asymptotically pivotal. That is, the bootstrap provides a better approximation to the distribution of an asymptotically pivotal statistic than does 'ordinary' asymptotic distribution theory. A statistic is asymptotically pivotal if its asymptotic distribution does not depend on unknown population parameters. All the familiar test statistics whose asymptotic distributions are standard normal or chi-square are asymptotically pivotal. Estimates of regression coefficients, standard errors, and other population parameters typically are not asymptotically pivotal. The bootstrap does not provide asymptotic refinements for statistics that are not asymptotically pivotal. Whenever possible, the bootstrap should be applied to asymptotically pivotal statistics as opposed to statistics that are not asymptotically pivotal.

The bootstrap's ability to provide asymptotic refinements has important practical consequences. Specifically, the bootstrap can be used to obtain estimates of finite-sample critical values for test statistics that are more accurate than critical values obtained from the asymptotic normal or chi-square approximations. The use of bootstrap-based critical values can greatly reduce the ERP of a test and ECP of a confidence interval.

The bootstrap provides asymptotic refinements because it provides a higher-order asymptotic approximation, called an Edgeworth approximation, to $G_n(\tau, F)$. Suppose that T_n is asymptotically distributed as N (0, 1), and let Φ denote the standard normal CDF. Then $G_n(\tau, F_n) - G_n(\tau, F) = O_p(n^{-1})$, whereas $G_n(\tau, F) - \Phi(\tau) = O(n^{-1/2})$. Thus, the error made by the bootstrap approximation to $G_n(\tau, F)$ converges to 0 more rapidly than does the error made by the asymptotic normal approximation. For $|T_n|$ or an asymptotic chi-square statistic, the error made by the bootstrap approximation is $O_p(n^{-3/2})$ whereas the error made by the asymptotic normal or chi-square approximation is $O(n^{-1})$. See Hall (1992) and Horowitz (2001) for details.

Rejection probabilities of tests and coverage probabilities of confidence intervals based on bootstrap critical values can be even more accurate. The ERPs of symmetrical tests and ECPs of symmetrical confidence intervals are $O(n^{-2})$ when the bootstrap is used to obtain the critical value, whereas they are $O(n^{-1})$ when the asymptotic normal or chi-square approximation is used. (A test based on an asymptotic chi-square statistic is symmetrical. So is a test that rejects the null hypothesis when $|T_n|$ exceeds the critical value, where T_n is asymptotically distributed as N (0, 1).) Thus, the ERPs and ECPs of symmetrical tests and confidence intervals converge to 0 much more rapidly with bootstrap-based critical values than with critical values based on the asymptotic normal or chi-square approximations. The practical consequence of this is that the bootstrap often achieves spectacular reductions in the numerical values of ERPs and ECPs. Section 3 provides two examples of this. Horowitz (1997, 2001) provides others.

With one-sided tests and confidence intervals, the ERP and ECP are usually $O(n^{-1})$ with bootstrap critical values and $O(n^{-1/2})$ with asymptotic chi-square or normal critical values. However, there are cases in which the ERP of a bootstrap-based test is $O(n^{-3/2})$ (Hall 1992; Davidson and MacKinnon 1999).

Examples

This section presents two examples that illustrate the bootstrap's ability to reduce the ERP of a test or the ECP of a confidence interval.

White's (1982) Information-Matrix (IM) Test

This is a specification test for parametric models estimated by maximum likelihood. The test statistic is asymptotically chi-square distributed, but the asymptotic distribution is a poor approximation to the finite-sample distribution.

Bootstrap, Table 1 Empirical rejection probabilities of nominal 0.05-level information-matrix tests of probit and tobit models

Rejection probability using					
N	Distr. of X	Asymp. critical values		Bootstrap crit. values	
		White	Chesh.-Lan	White	Chesh.-Lan.
Binary probit models					
50	N(0,1)	0.385	0.904	0.064	0.056
	U(− 2,2)	0.498	0.920	0.066	0.036
100	N(0,1)	0.589	0.848	0.053	0.059
	U(− 2,2)	0.632	0.875	0.058	0.056
Tobit models					
50	N(0,1)	0.112	0.575	0.083	0.047
	U(− 2,2)	0.128	0.737	0.051	0.059
100	N(0,1)	0.065	0.470	0.038	0.039
	U(− 2,2)	0.090	0.501	0.046	0.052

Source: Horowitz (1994)

Horowitz (1994) reports the results of Monte Carlo experiments that investigate the ERPs of the IM test with bootstrap critical values. Some of these results are summarized in Table 1, which gives the results of applying the Chesher (1983) and Lancaster (1984) form and White's (1982) original form of the test to Tobit and binary probit models. The results show that the ERPs are very large when critical values based on the asymptotic chi-square distribution are used. When bootstrap critical values are used, however, the ERPs are very small. The bootstrap essentially eliminates the differences between the true and nominal rejection probabilities of the two forms of the IM test.

Estimation of Covariance Structures

In estimation of covariance structures, the objective is to estimate the covariance matrix of a $k \times 1$ vector X subject to restrictions that reduce the number of unique, unknown elements to $r < k(k + 1)/2$. Estimates of the r unknown elements can be obtained by minimizing the weighted distance between sample moments and the estimated population moments. Weighting all sample moments equally produces the equally weighted minimum distance (EWMD) estimator, whereas choosing the weights to maximize asymptotic estimation efficiency produces the optimal minimum distance (OMD) estimator.

The OMD estimator has poor finite-sample performance in applications (Abowd and Card 1989). Horowitz (1998) reports the results of a Monte Carlo investigation of the ability of the bootstrap to reduce the ERPs of nominal 95 per cent symmetrical confidence intervals based on the OMD estimator. In each experiment, X has 10 components, and the sample size is $n = 500$. The j'th component of X, X_j ($j = 1, \ldots, 10$) is generated by $X_j = (Z_j + \rho Z_{j+1})/(1 + \rho^2)^{1/2}$, where $Z_1, \ldots, Z_{11}$ are i.i.d. random variables with means of 0 and variances of 1, and $\rho = 0.5$. The Z's are sampled from five different distributions depending on the experiment. It is assumed that ρ is known and that the components of X are known to be identically distributed and to follow MA(1) processes. The estimation problem is to infer the scalar parameter θ that is identified by the moment conditions $Var(X_j) = \theta$ ($j = 1, \ldots, 10$) and $Cov(X_j, X_{j-1}) = \rho\theta/(1 + \rho^2)$ ($j = 2, \ldots, 10$).

The results of the experiments are summarized in Table 2. The coverage probabilities of confidence intervals based on asymptotic critical values are far below the nominal value of 0.95 except in the experiment with uniform Z's. However, the use of bootstrap critical values greatly reduces the ERPs. In the experiments with normal, Student t, uniform, or exponential Z's, the bootstrap essentially eliminates the errors in the coverage probabilities of the confidence intervals.

Bootstrap, Table 2 Empirical coverage probabilities of nominal 95 per cent symmetrical confidence intervals based on the OMD estimator

Distr. of Z	Asymptotic critical value	Bootstrap critical value
Uniform	0.93	0.96
Normal	0.85	0.95
Student t with 10 d.f.	0.79	0.95
Exponential	0.54	0.96
Lognormal	0.03	0.91

Source: Horowitz (1998)

Acknowledgments I thank Federico Bugni for helpful comments. The preparation of this article was supported in part by NSF Grant SES-0352675.

Bibliography

Abowd, J.M., and D. Card. 1989. On the covariance of earnings and hours changes. *Econometrica* 57: 411–445.

Andrews, D.W.K. 1997. A conditional Kolmogorov test. *Econometrica* 65: 1097–1128.

Beran, R., and G.R. Ducharme. 1991. *Asymptotic theory for bootstrap methods in statistics*. Montréal: Les Publications CRM, Centre de Recherches Mathematiques, Université de Montréal.

Chesher, A. 1983. The information matrix test. *Economics Letters* 13: 45–48.

Davidson, R., and J.G. MacKinnon. 1999. The size distortion of bootstrap tests. *Econometric Theory* 15: 361–376.

Davison, A.C., and D.V. Hinkley. 1997. *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

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

Efron, B., and R.J. Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Hall, P. 1992. *The bootstrap and edgeworth expansion*. New York: Springer.

Hall, P. 1994. Methodology and theory for the bootstrap. In *Handbook of econometrics*, ed. R.F. Engle and D.F. McFadden, Vol. 4. Amsterdam: North-Holland.

Horowitz, J.L. 1994. Bootstrap-based critical values for the information matrix test. *Journal of Econometrics* 61: 395–411.

Horowitz, J.L. 1997. Bootstrap methods in econometrics: Theory and numerical performance. In *Advances in economics and econometrics: Theory and applications, seventh world congress*, ed. D.M. Kreps and K.F. Wallis, Vol. 3. Cambridge: Cambridge University Press.

Horowitz, J.L. 1998. Bootstrap methods for covariance structures. *Journal of Human Resources* 33: 39–61.

Horowitz, J.L. 2001. The bootstrap in econometrics. In *Handbook of econometrics*, ed. J.J. Heckman and E.E. Leamer, Vol. 5. Amsterdam: North-Holland.

Horowitz, J.L. 2003. The bootstrap in econometrics. *Statistical Science* 18: 211–218.

Horowitz, J.L., and C.F. Manski. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95: 77–84.

Lahiri, S.N. 2003. *Resampling methods for dependent data*. New York: Springer.

Lancaster, T. 1984. The covariance matrix of the information matrix test. *Econometrica* 52: 1051–1053.

Maddala, G.S., and J. Jeong. 1993. A perspective on application of bootstrap methods in econometrics. In *Handbook of statistics*, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod, Vol. 11. Amsterdam: North-Holland.

Mammen, E. 1992. *When does bootstrap work? Asymptotic results and simulations*. New York: Springer.

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

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

Politis, D.N., J.P. Romano, and M. Wolf. 1999. *Subsampling*. New York: Springer.

Shao, U., and D. Tu. 1995. *The jackknife and bootstrap*. New York: Springer.

Vinod, H.D. 1993. Bootstrap methods: Applications in econometrics. In *Handbook of statistics*, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod, Vol. 11. Amsterdam: North-Holland.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–26.

Borch, Karl H. (1919–1986)

Knut K. Aase

Keywords

Borch, K.; Game theory; General equilibrium; Insurance; Norwegian School of Economics and Business Administration (NHH); Reinsurance contracts; Risk

JEL Classifications

B31

Karl Borch was born in Sarpsborg, Norway, on 13 March 1919. He graduated with an MSc in actuarial mathematics at the University of Oslo in 1947, and a Ph.D. in 1962.

From 1947 he worked for UNESCO and OECD until in 1959 he started his academic career at the Norwegian School of Economics and Business Administration (NHH) in Bergen, where he was appointed professor of insurance in 1963, a position he held until his untimely death on 2 December 1986, only just before his retirement was due.

In *Who's Who in Economics* (1986, p. 103) he wrote: 'When in 1959 I got a research post which gave me almost complete freedom, as long as my work was relevant to insurance, I naturally set out to develop an economic theory of insurance.' That within a year he should have made a decisive step in that direction is amazing. What he did during these first years of his research career was to write the first of a long series of seminal papers, which were to put him on the map as one of the world's leading scholars in his field.

One important contribution of his papers in *Skandinavisk Aktuarietidskrift* (1960a) and *Econometrica* (1962) was to derive testable implications from the abstract model of general equilibrium with markets for contingent claims. In this way, he brought economic theory to bear on insurance problems, thereby opening up that field considerably; and he brought the experience of reinsurance contracts to bear on the interpretation of economic theory, thereby considerably enlivening that theory.

Practically his entire production was centred on the topic of uncertainty in economics. Many of his thoughts were formulated in his successful book *The Economics of Uncertainty* (1968a), also available in Spanish, German and Japanese. He gave the first graduate lectures at NHH, where he supervised many Ph.D. students.

He had more than 150 publications, among them three books (published in 1968, 1974 and 1990). The last one, *Economics of Insurance*, has also been translated into Chinese. Best known to actuaries is perhaps his pioneering work on Pareto-optimal risk exchanges in reinsurance (for example, 1960a). Borch also made many contributions to the application of game theory to insurance: in particular, he characterized the Nash bargaining solution of a reinsurance syndicate (1960b).

Borch served on many editorial boards, and he helped organize several key international conferences abroad and at NHH.

See Also

- ▶ General Equilibrium
- ▶ Insurance Mathematics
- ▶ Pareto Principle and Competing Principles
- ▶ Risk
- ▶ Risk Aversion
- ▶ Uncertainty

Selected Works

1960a. The safety loading of reinsurance premiums. *Skandinavisk Aktuarietidsskrift* 43: 163–84.

1960b. Reciprocal reinsurance treaties seen as a two-person cooperative game. *Skandinavisk Aktuarietidsskrift* 43: 29–58.

1962. Equilibrium in a reinsurance market. *Econometrica* 30: 424–444.

1967. The theory of risk. *Journal of the Royal Statistical Society*, Series B 29: 423–467.

1968a. *The economics of uncertainty.* Princeton, NJ: Princeton University Press.

1968b. General equilibrium in the economics of uncertainty. In *Risk and uncertainty*, ed. K. Borch and J. Mossin. London: Macmillan.

1969. A note on uncertainty and indifference curves. *Review of Economic Studies* 36: 1–4.

1974. *The mathematical theory of insurance*. Lexington, MA: D.C. Heath.

1985. A theory of insurance premiums. *The Geneva Papers on Risk and Insurance* 10: 192–208.

1986. Entry in *who's who in economics: A biographical dictionary of major economists, 1700–1986*, 2nd edn., ed. M. Blaug. Cambridge, MA: MIT Press.

1990. (With K. Aase and A. Sandmo, ed.) *Economics of insurance*. Advanced textbooks in economics 29. Amsterdam/ New York/ Oxford/ Tokyo: North-Holland.

Borda, Jean-Charles de (1733–1799)

Charles K. Rowley

Keywords

Black, D.; Borda, J.-C. de; Condorcet criterion; Condorcet, Marquis de; Laplace, Marquis de; Mathematical theory of elections; Method of marks; Public choice; Voting

JEL Classifications

B31

The second half of the 18th century in France was one of the outstanding epochs of scientific thought and witnessed significant attempts to carry the methods of rigorous and mathematical thought beyond the physical and into the realms of the human sciences. A brilliant start was made in political science by three French academicians, namely Borda, Condorcet and Laplace, with contributions which now play a central role in the literature of public choice. It is a salutary warning to those who view science as endlessly progressive to note that the contributions of these outstanding academicians were lost for two centuries until they were rediscovered in 1958 by Duncan Black.

Borda was the first of the three to develop a mathematical theory of elections shortly after becoming a member of the Academy of Sciences. Born in 1733 in Dax, near Bordeaux, Borda was successively an officer of cavalry, a naval captain, and a scholar of mathematical physics as well as an innovator in the field of scientific instruments. Newly elected to the Academy of Sciences, Borda read a paper entitled 'Sur la forme des elections' on 16 June 1770. Four members were charged to report on it, but failed to do so.

The Academy was not to consider elections again during the succeeding 14 years, until Borda again read a paper on elections in July 1784 following the favourable report by Bossut and Coulomb on Condorcet's manuscript, *Essai.* Borda's paper had been printed in the *Histoire de l' Academie Royale des Sciences* in 1781, three years prior to this reading. It was finally published in 1784. In essence, it reflected the content of his 1770 paper. Condorcet had become acquainted with Borda's contribution prior to writing his *Essai,* as a consequence of the strong oral tradition of the Academy. He acknowledged the powerful influence of Borda's ideas upon his own writings.

Borda was concerned that the single vote system of elections might select the wrong candidate. He illustrated by reference to a situation in which eight electors had candidate A as first preference, seven had candidate B, and six had candidate C. On the single vote, A would be elected, although the electors preferred B *or* C to A by a majority of 13 to 8. In essence, Borda was utilizing what later became known as the *Condorcet criterion,* though he failed to develop it himself. Instead, he attempted to remedy the defect of the single vote system by the method of marks, which he presented in two forms. Since one form is a special case of the other, only the more general form is here outlined.

The method of marks requires each elector to rank all the candidates by order of merit. The candidate is then allocated marks by reference to his ranking by each voter, for example, three marks for first place, two marks for second, and one mark for last in a three candidate election. The marks are then totalled across all elections. The candidate with the largest aggregate of marks is the winner.

To illustrate how the method of marks may provide a different result from that of the single vote, let us expand Borda's original example as outlined above into the form of Table 1.

In the Table 1 example, Candidate A would receive an aggregate of 39 marks, Candidate B receives an aggregate of 41 marks, and Candidate C receives an aggregate of 46 marks. Candidate C is the winner, reversing the single vote outcome.

Borda, Jean-Charles de (1733–1799), Table 1 Rank order of candidates by electors

A	A	B	B	C	C
B	C	A	C	A	B
C	B	C	A	B	A
1	7	1	6	1	5

The method of marks allows a role for preference intensities, albeit only on a strictly linear scale, within the electoral process. For this reason, it has been called a 'neo-utilitarian' approach (Sudgen 1981). The method is not strategy proof, since voters will tend to lower the ranking of the candidate most threatening to their preferred candidate to the lowest level, irrespective of their actual preferences. Borda himself clearly recognized this danger, but, in an age more honourable than our own, was merely moved to comment: 'My scheme is only intended for honest men.'

Borda's paper did not attempt to provide a comprehensive theory of elections. It failed to develop, though it implicitly embraced, the criterion of Condorcet. More important, it offered no real insight into the nature and/or the objectives of group decisions. It was, however, a significant first step in both directions. The method of marks is extremely effective if each elector genuinely desires to secure the election of 'that candidate who should be the most generally acceptable' (Black 1958). In reality, most electors desire to secure the election of their most favoured candidate. Herein lies the weakness of the method of marks.

Shortly after hearing Borda's paper in 1784, the Academy adopted his method in elections to its membership. The method of marks remained in use until 1800, when it was attacked by a new member, and soon afterwards, was modified. The new member in question was Napoleon Bonaparte.

See Also

- ▶ Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794)
- ▶ Social Choice
- ▶ Voting Paradoxes

Selected work

1781. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royal des Sciences,* Paris.

Bibliography

Black, D. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

de Condorcet, Marquis. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.

de Laplace, Marquis. 1812. Leçons de mathématique données à l'Ecole Normale en 1885. *Journal de l'Ecole Polytechnique.*

Lacroix, S.E. 1800. *Eloge historique de Jean-Charles Borda*. Paris: Imprimerie de R. Jacquin.

Mascart, J. 1919. *La vie et les travaux du Chevalier Jean-Charles de Borda (1733–1799): épisodes de la vie scientifique au XVIII siécle*. Lyon: Imprimerie A. Rey.

Sudgen, R. 1981. *The political economy of public choice: An introduction to welfare economics*. Oxford: Martin Robertson.

Border Effects

John H. Rogers

Abstract

International finance and trade economists have traditionally focused on the behaviour of cross-country prices and factor returns and the flow of goods and capital across nations. Studying these same variables across locations within countries provides a baseline for measuring the influence of the border. The 'border effect' is the difference between international and intra-national magnitudes. Large border effects were initially found in consumer goods prices and trade volumes. Subsequent studies have examined robustness and looked for explanations of the border effect.

Keywords

Border effects; Distance; Exchange rate volatility; Gravity models; Internal trade; International trade barriers; Market integration; Price

volatility; Productivity shocks; Purchasing power parity; Sticky prices; Trade flows; Transaction costs; Transportation costs; United States–Canada Free Trade Agreement

JEL Classifications
F3

International finance and trade economists have traditionally focused on the behaviour of cross-country prices and factor returns and the flow of goods and capital across nations. Studying these same variables across locations *within countries* provides a baseline for measuring the influence of the border. The 'border effect' is, to speak loosely, the difference between international and intra-national magnitudes. Large border effects were initially found in consumer goods prices and trade volumes (Engel and Rogers 1996; McCallum 1995) in data from the United States and Canada. Subsequent studies have examined robustness and looked for explanations of the border effect, often through extensions to other countries' data-sets.

The starting point of Engel and Rogers (1996) is a fundamental proposition of economic theory: in the absence of transaction costs, identical goods must sell for the same price. Prices will fail to equalize when there are barriers, natural ones or man-made, to the free movement of goods. There are several reasons to expect that national borders would give rise to such barriers.

Engel and Rogers (1996) examine the behaviour of prices of 14 categories of consumer goods and services in 14 US cities and nine Canadian cities during the period 1978–94. They measure the border effect by comparing the extent to which prices of a particular category of goods fluctuate across cities intra-nationally with price fluctuations for city pairs that lie across the border. With q_{ij} defined as the log of the price of some good in city i relative to its price in city j, let $V(q_{ij})$ be a measure of relative price volatility over the sample time period. Engel and Rogers relate this to various explanatory variables including distance between cities and a 'border dummy' for whether the cities lie in different countries. They run regressions of the form:

$$V(q_{ij}) = \beta_1 d_{ij} + \beta_2 B_{ij} + \sum_{k=1.m} \lambda_k D_k, \qquad (1)$$

where d_{ij} is the log of the distance between cities i and j; B_{ij} is a dummy variable equal to 1 if cities i and j are in different countries; and D_k are dummy variables for each city. Engel and Rogers (1996) consistently find that β_2 is positive, highly statistically significant, and large in magnitude. The coefficient on distance, β_1, is usually positive and significant.

McCallum (1995) estimates the effect of the border on trade flows between Canadian provinces and US states. McCallum's data-set includes imports and exports for all pairs of Canadian provinces, as well as imports and exports between each of the ten provinces and each of the 50 US states. The data are from 1988. McCallum uses a traditional gravity model, positing that trade is a function of the distance between trading partners and their individual economic sizes, measured by gross domestic product. (See Anderson 1979, for model development, and Rose 2000, for a noteworthy application.) McCallum augments the standard gravity model with a dummy variable equal to 1 for pairs of Canadian provinces.

The coefficient on McCallum's inter-provincial trade dummy variable is estimated to be positive and highly statistically significant. The point estimate implies that, other things equal, trade between two Canadian provinces is more than 20 times larger than trade between a province and a US state.

Anderson and van Wincoop (2003) are critical of the gravity equations employed in the border effects papers on trade flows. They argue that these equations suffer from omitted variables bias (requiring that a 'multi-lateral resistance' term be added) and incorrect comparative statics analysis. Anderson and van Wincoop develop a methodology that allows them to get around these shortcomings. Taking up McCallum's exercise using data for 1993, these authors show that the border effect on trade flows is, although still large, considerably smaller than calculated by McCallum.

Engel and Rogers (1996) suggest several reasons why the border should matter. First, there might be direct costs to crossing the border such as tariffs and other trade restrictions. Alternatively, markups might differ across locations and vary with exchange rate changes. Markets for non-traded inputs (wages, marketing services) might be more highly integrated on a national basis than in two places separated by a border. Or productivity shocks might be more similar for city pairs that lie within a country than for cross-border pairs. Finally, Engel and Rogers consider a sticky-price explanation. Goods sold in the United States may be sticky in US dollar terms while goods sold in Canada are sticky in terms of Canadian dollars. A highly variable nominal exchange rate could then give rise to a large, positive value of β_2 because cross-border relative prices would fluctuate along with the nominal exchange rate while relative prices within countries remained fairly stable. Although Engel and Rogers do not conduct an exhaustive examination of different factors, they conclude, 'Sticky prices appear to be one explanation but probably do not explain most of the border effect' (1996, p. 1112). (The Engel–Rogers work has an intellectual predecessor in Mussa 1986, who noted that CPI-based real exchange rates are more variable for Toronto versus Chicago, Vancouver versus Chicago, Toronto versus Los Angeles, and Vancouver versus Los Angeles, than for Toronto versus Vancouver and Chicago versus Los Angeles under floating exchange rates. Mussa attributed this to sticky prices.)

Using updated data, Engel and Rogers (2000) examine the stability of the border effect around the United States–Canada Free Trade Agreement. They find little evidence of a change across several break dates corresponding with the signing or implementation of the agreement.

Subsequent studies have examined different data-sets and attempted to understand the dynamics of the border effect. Parsley and Wei (2001) examine data from 96 US and Japanese cities from 1976 to 1997. They ask two related questions. First, is there any evidence that the Japan–US 'border' narrows over time? Second, is there evidence linking the evolution of the border effect with plausible economic candidates (for example, the unit cost of international transportation)? They show that the simple average of good-level real exchange rates tracks the nominal exchange rate closely, providing strong evidence of sticky prices in local currencies. They find evidence that the border effect between Japan and the United States declines over time. Furthermore, distance, shipping costs, and exchange rate variability collectively explain a substantial portion of the border effect.

Engel and Rogers (2001) use consumer price data from European cities in 11 countries from 1981 to 1997 to explore deviations from short-run purchasing power parity (PPP) across several national borders. The European data-set has many advantages over that consisting of observations from US and Canadian cities only. In the latter, there is no distinction between the border dummy and a measure of nominal exchange rate variability, since all cross-border pairs have the same nominal exchange rate. With the European data-set, Engel and Rogers are able to include both a border dummy variable (unity for city pairs lying across the border) and a measure of nominal exchange rate variability. This allows a distinction between the role of sticky local currency pricing and the various other 'real' barriers to market integration. The authors find that, even with nominal exchange rate variability taken into account, distance between cities and the border continue to have positive and significant effects on real exchange rate variability. However, these effects are smaller than the local currency pricing effect.

Gorodnichenko and Tesar (2005) re-examine the Engel–Rogers and Parsley–Wei papers. They run the same regression as the earlier papers but propose a different measure of the border effect. To understand their measure, let γ_U be the average relative price variance for city pairs within the United States; γ_C be the average for pairs within Canada; and β be the average relative price variance for cross-border city pairs (after controlling for distance). Engel and Rogers (1996) measure the border effect as $\beta - 0{:}5(\gamma_U + \gamma_C)$. Gorodnichenko and Tesar propose the 'conservative' measure: β -max(γ_U; γ_C). Since γ_U is not very different from β (a feature of the data noted by

Engel and Rogers), the border effect is small when measured in the conservative way. Under the Gorodnischenko–Tesar scheme there are two border effects, one for a Canadian crossing into the US market and the other for an American crossing into Canada. In this case one is quite large, the other relatively small. Engel and Rogers measure the border effect as the average of the two (as do Parsley and Wei for the US–Japan data). Gorodnichenko and Tesar use the smaller of the two.

A large body of literature has expanded upon McCallum's (1995) findings. As with the literature that followed Engel and Rogers (1996), many have analysed different data-sets, especially from other countries. Examples include Helliwell (1996, 1998), Wei (1996), Anderson and Smith (1999), Yi (2003), Wolf (2000), Hillberry and Hummels (2003), and Evans (2003). One important issue highlighted by these papers is the need for accurate measures of 'internal trade', that is, the amount that countries trade with themselves. This literature is exhaustively surveyed by Anderson and van Wincoop (2004).

Progress in explaining the border effect on trade flows has been made by decomposing total international trade barriers into barriers associated with geographic factors such as distance and barriers due to national borders. According to Anderson and van Wincoop (2004, Table 7), estimates from several papers using different data-sets (Wei 1996; Eaton and Kortum 2002; Evans 2003; Anderson and van Wincoop 2003) put the tariff equivalent cost of total international trade barriers at around 40–80 per cent. Anderson and van Wincoop categorize further investigation of the trade barriers associated with national borders as attempts to quantify the effects due to (*a*) language barriers, (*b*) use of different currencies, (*c*) information barriers, (*d*) contracting costs and security, and (*e*) policy barriers. To summarize the results from this literature, these authors suggest very rough calculations of an eight per cent policy-related barrier, a seven per cent language barrier, a fourteen per cent currency barrier, a six per cent information cost barrier, and a three per cent security barrier, well within the range of 25–50 per cent for overall border barriers reported by different authors for OECD countries.

See Also

- ▶ Exchange Rate Volatility
- ▶ Gravity Models
- ▶ Price Dispersion
- ▶ Trade Costs

Bibliography

Anderson, J. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69: 106–116.

Anderson, J., and E. van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93: 170–192.

Anderson, J., and E. van Wincoop. 2004. Trade costs. *Journal of Economic Literature* 42: 691–751.

Anderson, M., and S. Smith. 1999. Do national borders really matter? Canada–U.S. regional trade reconsidered. *Review of International Economics* 7: 219–227.

Eaton, J., and S. Kortum. 2002. Technology, geography and trade. *Econometrica* 70: 1741–1779.

Engel, C., and J. Rogers. 1996. How wide is the border? *American Economic Review* 86: 1112–1125.

Engel, C., and Rogers. 2000. Relative price volatility: What role does the border play? In *Intranational macroeconomics*, ed. G. Hess and E. van Wincoop. Cambridge: Cambridge University Press.

Engel, C., and J. Rogers. 2001. Deviations from purchasing power parity: Causes and welfare costs. *Journal of International Economics* 55: 29–57.

Evans, C. 2003. The economic significance of national border effects. *American Economic Review* 93: 1291–1312.

Gorodnichenko, Y, and L. Tesar. 2005. *A re-examination of the border effect*. Working paper no. 11706. Cambridge, MA: NBER.

Helliwell, J. 1996. Do national boundaries matter for Quebec's trade? *Canadian Journal of Economics* 29: 507–522.

Helliwell, J. 1998. *How much do national borders matter?* Washington, DC: Brookings Institution.

Hillberry, R., and D. Hummels. 2003. Intranational home bias: Some explanations. *The Review of Economics and Statistics* 85: 1089–1092.

McCallum, J. 1995. National borders matter: Canada–U.S. regional trade patterns. *American Economic Review* 85: 615–623.

Mussa, M. 1986. Nominal exchange rate regimes and the behavior of real exchange rates: Evidence and implications. *Carnegie-Rochester Conference Series on Public Policy* 25: 117–214.

Parsley, D., and S.J. Wei. 2001. Explaining the border effect: The role of exchange rate variability, shipping costs and geography. *Journal of International Economics* 55: 87–105.

Rose, A. 2000. One money, one market: Estimating the effect of common currencies on trade. *Economic Policy* 30: 7–45.

Wei, S.J. 1996. *Intra-national versus inter–national trade: How stubborn are nations in global integration?* Working paper no. 5531. Cambridge, MA: NBER.

Wolf, H. 2000. (Why) do borders matter for trade? In *Intranational Macroeconomics*, ed. G. Hess and E. van Wincoop. Cambridge: Cambridge University Press.

Yi, K.M. 2003. *A simple explanation for the border effect*. Working paper, Federal Reserve Bank of New York.

Bortkiewicz, Ladislaus von (1868–1931)

Luca Meldolesi

'By far the most eminent German statistician since Lexis' (Schumpeter 1932, p. 338), Bortkiewicz was born in St Petersburg into a family of Polish origin and educated in a Russian cultural environment (the University of St Petersburg included). Later, encouraged by W. Lexis and G.F. Knapp, he studied at the University of Strasbourg, where for two years he also taught, as *Privat-Dozent*, accident insurance and theoretical statistics. Back to St Petersburg, he worked from 1899 to 1901 at the Aleksandr Liceo – an elite secondary school of Russian *étatisme*. Then he was appointed 'extraordinary' (i.e. assistant) professor of economics and statistics at the University of Berlin, where he taught for 30 years, receiving his full professorship in 1920.

Bortkiewicz's work covers a wide range of subjects on statistics, economics, mathematics, even physics, and is scattered in a large number of publications. Bortkiewicz is considered one of the few great scholars of his time in the field of statistical methodology. His 'law of small numbers' or of 'rare events' (*Das Gesetz der Kleinen Zahlen* 1898a) won great scientific attention – and unleashed an animated polemic in *Giornale degli Economisti* (1907–9) – particularly through the almost miraculous application of this law to the 280 Prussian soldiers killed by the kicks of their horses in the period 1875–94. An incomplete list of Bortkiewicz's writings published by Oskar Anderson in 1931 includes 54 entries – books, essays, notes – on 'theoretical statistics and calculus of probability'. Of these, Schumpeter pointed out to the economist a book (*Die Iterationen* 1917) and papers on the measure of income inequality (1930), the quadrature of empirical curves (1926), homogeneity and stability in statistics (1918), variability under the Gaussian Law (1912), the property common to all laws of error (1923b) and the succession in time of chance events (1911).

As for economics, Bortkiewicz's writings – at least 24 papers – range from the theory of value of monetary theory and policy. (Contributions in the latter focus on the gold standard, banking credit, the velocity of circulation, and index numbers (1924).) As is known, some papers on the theory of value have particularly attracted an enduring attention. In 1949 Paul M. Sweezy published the English translation of an article on Marx (Bortkiewicz 1907). In 1952, two sections of Bortkiewicz's 1906–7 long essay, 'Wertrechnung und Preisrechnung im Marxschen System', were translated in *International Economic Papers*. Finally, in 1971, a group of essays on the economic theories of Marx, Böhm-Bawerk, Walras and Pareto, was collected and published in Italian.

Bortkiewicz was essentially a critic. According to Oskar Anderson (1931) his analytic mind was extraordinarily acute, cold and merciless with mistakes and sloppy arguments, so that he was universally considered a stern and even quick-tempered judge, whose review articles nobody could overlook. To make his intellectual machine work he needed an external stimulus, often provided by scientific contributions of well-known authors. He entered into them, elaborated on them and sometimes confuted them.

These peculiarities are at work in his famous criticism of Böhm-Bawerk's theory of the origin of the interest on capital ('Der Kardinalfehler der Böhm-Bawerkschen Zinstheorie' 1906a). Bortkiewicz believed that the theses put forward

by the 'theory of productivity' on this subject had been definitely confuted by Böhm-Bawerk, but that the alternative explanations suggested by him were also objectionable. According to Böhm-Bawerk's main argument, longer methods of production are technically more productive than shorter ones, so that present capital goods provide us with quantities of consumption goods greater than future capitals: this is the source of the interest of capital.

However, objects Bortkiewicz, a maximum level of production for each given capital invested always exists: because of physical reasons – if nothing else. Therefore, if we compare two investments, started at different times but equal in amount and composition, each of them will produce the same output, but at a temporal distance corresponding to the initial interval. Hence Böhm-Bawerk's superiority of present capital goods over future ones turns out to be a simple time span, which, in itself, is unable to explain the origin of interest.

At this point Bortkiewicz focuses his attention on a different explanation proposed by Böhm-Bawerk (and others): the scarcity of capital. The latter, Bortkiewicz maintains, can only be temporary and due to mistaken foresight. Since capital, according to Böhm-Bawerk, is nothing but an 'intermediate product', the working of the market mechanism will ease and eventually cancel out the shortages of the different capitals (vis-à-vis the workers) in the different lines of production.

On the other hand, this criticism of Böhm-Bawerk finds its *pendant* in Bortkiewicz's appreciation of Marx's theses on the origin of profit (and interest). 'Wertrechnung', Bortkiewicz's main article on Marx, was published shortly after 'Die Kardinalfehler . . .' and is part of the same line of thinking. (Later, Bortkiewicz also came back to the problem in an essay significantly titled 'Böhm-Bawerk's main work in his relation to the socialist theory of the interest on capital' (1923a)).

'Wertrechnung' is divided into three parts. The first is dedicated to a long survey of opposers, followers and independent observers of Marx's conception of value and price. The author places himself among the 'mediators', in Lexis' footsteps (and recalls that Lexis' criticism (1885) had been favourably taken up by Engels in his preface to the third volume of *Das Kapital*).

The second part contains the well-known determination of prices and profit rate based on equations originally put forward by the Russian economist Dmitriev (1904) in his work on Ricardo. This solution that bears out many of Ricardo's propositions can be usefully compared with a second one published by the author at the same time (July 1907) in 'Zur Berichtigung'.

Here, by taking into account Tugan-Baranovsky's contribution on the subject (1905), Bortkiewicz actually develops a suggestion made by Marx himself, according to which the values of inputs should be transformed into prices as well as the values of outputs.

The two solutions are shown to stem from different but connected ways of analysing (circulating) capital and therefore to be part of a single theoretical structure: they can be generalized, and eventually come to the same results (Garegnani 1960; Meldolesi 1971). Given the wage rate and the unit of account, they determine simultaneously prices and the profit rate, which in turn depend on the processes directly and indirectly used in the production of the wage commodities alone. However, from all this – and the connected results on the falling rate of profit, absolute rent (1910–11), and so on – an 'objectivist' stand should not be inferred. Bortkiewicz believed that both objective and subjective influences on prices should be recognized and that his cost equations could be inserted into the wider setting of general equilibrium analysis (1890, 1898b, 1906, 1907, 1921) – a hypothesis that, after the debate on Sraffa's 'reswitching of techniques' (1960, vol. III), is by now rather discredited.

The third part of 'Wertrechnung' discusses the theory of profit and culminates, as one might expect, on the *origin* of profit (and interest). In comparison to Ricardo, Bortkiewicz suggests, Marx had a fortunate inspiration in building a mode in which profit (as surplus-value) exists while commodities exchange according to values alone. For, in such a system, it is obvious that profit can neither come about through raising prices in the exchange, nor can it be the reward for 'capital productive services'. In other words,

starting from values, Marx has defined in a clear and more significant way the theory of exploitation (or of deduction, as Bortkiewicz calls it with a neutral terminology) and has succeeded in making confusion on the matter impossible.

See Also

- ▶ Transformation Problem
- ▶ Value and Price

Selected Works

1890. *Eléments d'économie politique pure* de Léon Walras. *Revue d'économie politique*.

1898a. *Das Gesetz der Kleinen Zahlen*. Leipzig.

1898b. Die Greznutzentheorie als Grundlage einer ultraliberalen Wirtschaftspolitik. *Schmollers Jahrbuch* 22.

1906a. Der Kardinalfehler der Böhm-Bawerkschen Zinstheorie. *Schmollers Jahrbuch* 30.

1906b. Wertrechnung und Preisrechnung im Marxschen System. *Archiv für Sozialwissenschaft und Sozialpolitik* 23 and 25.

1907. Zur Berichtigung der grundlegenden theoretischen Konstruktion von Marx im dritten Band des 'Kapital', *Jahrbücher für Nationalökonomie* 24.

1910–11. Die Rodbertus'sche Grundrententheorie und die Marx'sche Lehre von der absoluten Grundrente. *Archiv für Geschichte des Soziolismus und der Arbeiterbewegung* 1.

1911. Sterbeziffern und Frauenüberschuss. *Bulletin de l'Institut International de Statistique*.

1917. *Die Iterationen. Ein Beitrag zur Wahrscheinlichkeitstheorie*. Berlin.

1918. Homogeneitäl und Stabilitäl in der Statistik. *Skandinavisk Aktuarietidskrift*.

1921. Objektivismus und Subjektivismus in der Werttheorie. In *Nationalekonomiska Studier till Knut Wicksell.* Stockholm.

1922. Die Variabilitätsbreite beim Gausschen Fehlergesetz. *Nordisk Statistisk Tidskrift* 1.

1923a. Böhm-Bawerks Hauptwerk in seinem Verhältnis zur sozialstatistischen Theorie des Kapitalzinses. *Archiv für Geschichte des Sozialismus und der Arbeiterbewegung*.

1923b. Über eine verschiedenen Fehlergesetzen gemeinsame Eigenschaft. In *Sitzungsberichte der Berliner mathematischenGesellschaft*. Berlin.

1924. Zweck und Struktur einer Preisindexzhl. *Nordisk Statistisk Tidskrift* 2 and 3.

1926. Über die Quadratur empirischer Kurven. *Skandinavisk Aktuarietidskrift*.

1930. Die Disparitätsmasse der Einkommensstatistik. In *XIX Session de l'Institut International de Statistique, Tokio 1930*. L'Aja.

1949. Appendix to E. von Böhm-Bawerk. *Karl Marx and the close of his system*, ed. P.M. Sweezy. New York: A.M. Kelley.

1952. Value and price in the Marxian system. *International Economic Papers* 2.

1971. *La teoria economica di Marx ed altri saggi su Böhm-Bawerk, Walras e Pareto*. ed. L. Meldolesi. Turin: Einaudi.

References

Anderson, O. 1931. Ladislaus v. Bortkiewicz [Obituary]. *Zeitschrift für Nationalökonomie* 3/2.

Dmitriev, V.K. 1904. *Economic essays on value, competition and utility.* Cambridge: Cambridge University Press. 1974.

Garegnani, P. 1960. *Il capitale nelle teorie della distribuzione*. Milan: Giuffrè.

Garegnani, P., et al. 1981. In *Valori e prezzi nella teoria di Marx*, ed. R. Panizza and S. Vicarelli. Turin: Einaudi.

Lexis, W. 1885. Die Marx'sche Kapitaltheorie. *Jarbücher für Nationalökonomie* 11.

Lexis, W. 1896. The concluding volume of Marx's Capital. *Quarterly Journal of Economics* 10.

Schumpeter, J.A. 1932. Obituary: Ladislaus von Bortkiewicz. *Economic Journal* 42: 338–340.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Sweezy, P.M. 1949. Editor's introduction. In *The close of his system*, ed. E. von Böhm-Bawerk, Karl Marx, and P.M. Sweezy. New York: A.M Kelley.

Tugan-Baranovsky, M. 1905. *Theoretische Grundlagen des Marxismus*. Leipzig: Duncker & Humblot.

von Meldolesi, L. 1971. Il contributo di Bortkiewicz alla teoria del valore, della distribuzione e dell'origine del profitto. In *La teoria economica di Marx ed altri saggi su Böhm-Bawerk, Walras e Pareto*, ed. L. Bortkiewicz. Turin: Einaudi.

Boulding, Kenneth Ewart (Born 1910)

Anatol Rapoport

Boulding was born on 18 January 1910 in Liverpool, England, and educated at Oxford and the University of Chicago. He has lived in the United States since 1937, teaching at Colgate, Fisk, Iowa State and McGill Universities, the University of Michigan and the University of Colorado. He was president of several learned societies including the American Economic Association and the American Association for the Advancement of Science.

The steadfast purpose that Boulding pursued in his work has been integration of knowledge. Instead of following the endlessly ramifying paths of specialized research in his chosen discipline, he sought to reach out from his 'home base' in economics to knowledge generated in other fields and, above all, to establish a leverage for deriving common vocabularies, conceptual frameworks, and methods.

This drive toward integration marks all of Boulding's contributions to economics. A typical example is the use of demographic models to describe macroeconomic aggregates. The sizes of biological populations are, of course, determined by birth rates and death rates. But these depend significantly on the age structure of the populations. Boulding conceives the aggregates of physical capital as a population of items, each characterized by an age. Production is analogous to births, consumption to deaths. Surely, the rates depend on the age structure of the 'population', as was vividly demonstrated in the post-war boom in the US automobile industry, when the population of automobiles was old (and hence had a high 'death rate') and by the eventual slump, which could have been predicted as a consequence of the same population becoming predominantly 'young'. The emphasis on 'structure' of aggregates marks also Boulding's treatment of income, the levels of prices and wages, price flexibility, etc.

To the extent that Boulding can be said to subscribe to any economic school of thought, he can be regarded as a Keynesian. His contribution in this direction has been in the macroeconomic theory of profits, which relates profits both to net investment and to distributions out of profits, what Keynes called the 'widow's cruse theory' (referring to the biblical legend of the cruse that never ran dry). While acknowledging an 'enormous debt to Keynes's brilliance of insight and imaginative sweep', Boulding points to a number of weaknesses in Keynesian macroeconomics, in particular failure to distinguish between exchanges, on the one hand, and the processes of production and consumption, on the other. In *A Reconstruction of Economics* Boulding developed separate theories of the two processes. In the same book he offered what he himself regards as, perhaps, the most original and controversial attempt to correct a weakness of Keynesian theory. The central idea is based on a generalization from the context of microeconomics to macroeconomics of the gross growth in the value of net worth.

Grant economics, that is, the theory of one-way transfers (in contrast to exchanges) is a field that Boulding helped to found. Subsidies, philanthropy and welfare clearly fall within the scope of this field. But it may well be extended to taxation or generally to any transactions involving transfers difficult to define as exchanges, the prime concern of mainstream economics.

In 'evolutionary economics', as in 'demography of aggregates', Boulding again draws upon the conceptual repertoire of biological science. Economics is seen as an evolving ecosystem, following the general principles of mutation and selection. Mutation is interpreted as new ideas, new knowledge, modified, of course, by monopoly, government policy, etc. 'Know how'. plays the role of the fundamental genetic factor, analogous to the seat of biological heredity, directing the development of the units of the ecosystem (analogues of organisms) whose interactions, in turn, shape the evolution of the system.

Recognition of Boulding's stature in the academic world has been lavish. Besides

professorships in six universities and presidencies of several learned societies, he has held visiting research and teaching positions in about twenty institutions all over the world. He has been the recipient of at least ten honorary degrees and as many medals, awards, and prizes.

In contrast, Boulding's profession (as an 'establishment') has exhibited a marked coolness toward his work. In pursuing his commitment, Boulding abandoned the safety of established theoretical frameworks and conceptual schemes of his discipline. In particular, his major work, *A Reconstruction of Economics* (1950), which was, perhaps, meant to introduce new paradigms in the development of economic theory, met with a mixed response. There was no lack of appreciation of Boulding's originality and felicitous insights, the outstanding traits of his writings. The soundness of his specific contributions, however, was at times questioned. William Vickrey, in his review of *A Reconstruction of Economics* wrote:

> The most interesting and suggestive, but perhaps precarious section is the last, in which Boulding carries the analysis of macroeconomic identities to new and perhaps extravagant lengths. The new superstructure, though it leads to very interesting and even startling conclusions, depends, in many crucial spots, on precisely the kind of structural stability of relationship, and absence of unanalyzed side effects that Boulding has been at pains to warn us of in the preceding section. (*American Economic Review* 1951, pp. 671–6)

Not only the content but also the style of Boulding's rich output (about forty books and a thousand articles) must have contributed to widening the gulf between him and the economic 'establishment'. Integration is accomplished in consequence of seeing unity in diversity. Accordingly, analogy plays a prominent role in Boulding's thought, the sort of analogy that serves as the mortar of a general theory of systems, where structural similarities connecting situations of widely differing content are at the focus of attention. (Boulding was a co-founder of the Society for General Systems Research.)

Analogies occupy positions on a spectrum of rigour. At one end are mathematical isomorphisms, providing the most solid basis for unified theories of widely different phenomena. At the other end are the metaphors of poetry, triggering at times exhilarating insights but not guaranteeing any degree of objective validity or theoretical leverage. Boulding travels freely over that spectrum. In consequence, his style is wholly devoid of the dullness traditionally expected in works with a claim to scientific rigour or scholarly erudition. It seems that Boulding's attraction to what is interesting and paradoxical and his undisguised delight in iconoclasm, as well as the paucity of his references to the work of other economists ('It is easier to think it up than look it up'), contributed to the estrangement between him and his profession.

Boulding has no compunction against stating a profound principle as a quip, for example, 'Everything that exists is possible' (primacy of empirical evidence over doctrinaire conclusions), 'Things are the way they are, because they got that way' (commitment to the evolutionary point of view). There are gems to be found in Boulding's delightful jingles: 'That is reckoned wisdom which/ Describes the scratch but not the itch' (a barb aimed at behaviourist dogma).

Boundaries between devotion to truth and devotion to values have no more meaning for Boulding than those between instruction and entertainment. He recognizes the stimulus that led many scientists to insist on a hermetic separation between 'what is' and 'what ought to be'. For instance, the superiority of 'price equilibrium' over a 'just price' as a fertile theoretical construct of economics is not disputed. But this emancipation from externally imposed morality has freed science in Boulding's estimation to develop its own system of values, apparent to anyone who, like Boulding, sees science not as an agglomeration of facts or techniques, not even as a system of theories but as an ongoing human enterprise, a passionate search for wisdom. Like Socrates, Boulding identifies wisdom with virtue. It is, perhaps, this insistence on the fundamental morality of science and of economics in particular that was the most important factor creating a distance between Boulding and the economic establishment.

Brought up as a Methodist and eventually becoming a Quaker, Boulding has remained a deeply religious person. For him Christianity is inseparable from pacifism. Rejection of war as an institution and of violence in all its manifestations is a cardinal principle in his political orientation. He has provided some outstanding leadership in the American peace movement, particularly during the turbulent years of the Vietnam war. He was a co-founder and director of the Center for Research in Conflict Resolution at the University of Michigan. This absorbing involvement in peace issues is reflected in several of his works, for example *The Economics of Peace* (1945), *Conflict and Defense* (1962), *Disarmament and the Economy* (1963).

In sum, Boulding is an economist who under pressure of intellectual curiosity and a devotion to freedom, justice, and progress (for which he has offered quite respectable operational definitions) has turned into a philosopher, be it noted, a scientifically literate one. The full flavour of his creative thought can be savoured in *The Meaning of the Twentieth Century* (1964), *Beyond Economics* (1968), and *The Image* (1956). The latter book, dictated in eleven days, was the 'product' of Boulding's sojourn at the Center for Advanced Study in the Behavioural Sciences in Palo Alto, California. There he met many of his contemporaries, who, he says, had a profound influence on his thinking.

Selected Works

1941. *Economic analysis*. New York: Harper. Rev. edn, Harper, 1948; 3rd edn, New York: Harper, 1955; 4th edn, New York: Harper & Row, 1966.
1945. *The economics of peace*. Englewood Cliffs, NJ: Prentice Hall; reissued, New York: Books for Libraries Press, 1972.
1950. *A reconstruction of economics*. New York: Wiley; reissued, Science Editions, 1962.
1953. *The organizational revolution: A study in the ethics of economic organizations*. New York: Harper, 1953; reissued, New York: Greenwood Press, 1984.
1956. *The image; knowledge in life and society*. Ann Arbor: University of Michigan Press.
1958a. *The skills of the economist*. New York, Toronto: Clarke, Irwin.
1958b. *Principles of economic policy*. Englewood Cliffs, NJ: Prentice Hall.
1962. *Conflict and defense: A general theory*. New York: Harper.
1963. (With E. Benoit, ed.) *Disarmament and the economy*. New York: Harper & Row; reissued, New York: Greenwood Press, 1978.
1964. *The meaning of the twentieth century: The great transition*. New York: Harper & Row.
1968. *Beyond economics: Essays on society, religion, and ethics*. Ann Arbor: University of Michigan Press.
1970. *Economics as a science*. New York: McGraw-Hill.
1973. *The economy of love and fear: A preface to grants economics*. Belmont, California: Wadsworth.
1978. (With T.F. Wilson, eds.) *Redistribution through the financial system: The grants economics of money and credit*. New York: Praeger.
1981. *Evolutionary economics*. London: Sage Publications.
1985. *Human betterment*. London: Sage Publications.

Bouniatian, Mentor (1877–1969)

Mauro Boianovsky

Abstract

Bouniatian argued that productive forces cannot be transferred to the future just through the accumulation of capital goods; the choice of production methods determines an equilibrium relation between aggregate consumption and the capital stock. Economic fluctuations are explained by both the increase of the proportion of income saved when output is growing and the period of time necessary for the

production of capital goods. The temporary separation between investment and consumption decisions is reflected in a more than proportional increase of capital goods. Changes in the marginal utility of consumption and capital goods and a generalization of the old King's law explain changes in the price level.

Keywords

Acceleration principle; Aftalion, A; Aggregate consumption; Bouniatian, M; Business cycles; Capital accumulation; Depreciation; Exogenous and endogenous business cycle theories; Great Depression; Hobson, J. A; Keynes, J. M; King's law; Lauderdale, Eighth Earl of; Mitchell, W. C; Mummery, A. F; Overproduction; Quantity theory of money; Saving–investment equality; Sticky wages; Subjective theory of value; Tugan- Baranovsky, M. I; Underconsumptionism; Weber–Fechner law

JEL Classifications

B31

Bouniatian was born in Ervian (Armenia) on 22 January 1877, and died on 31 January 1969 in Montmorency (near Paris). He received a D. Sc. from the University of Munich in 1903, and then taught at the University of Moscow and at the Polytechnical Institute of Tiflis (Georgia). From 1916 to 1919 he was manager of the Merchants Bank of Tiflis. After emigrating to France in 1920 as a political refugee, Bouniatian served on the faculty of law of the University of Paris from 1925 to 1940. He later became director of the Office of Armenian Refugees (a public service of the French ministry of foreign affairs) from 1945 to 1952.

Bouniatian's main contribution to economics is contained in his *Studien zur Theorie und Geschichte der Wirtschaftskrisen,* published in two volumes dated 1908. (Bouniatian often pointed out that the book actually came out in October 1907; the date issue was important to his claim that many of his ideas were later incorporated in Albert Aftalion's better-known articles and books. In fact, the list of books received in the February 1908 issue of the *Journal of Political Economy* gives 1907 as the date of publication.) In the first volume Bouniatian put forward a theory of the business cycle based on an original combination of elements from the underconsumption tradition and the then new accelerator concept, plus a novel explanation of changes in the price level. The volume was later revised and translated into Russian (1915) and French (1922; 1930). English expositions can be found in two articles by Bouniatian (1928, 1934). The second volume of the 1908 set is a detailed historical investigation of economic crises in England in the two centuries from 1640 to 1840, which provided the empirical basis for the theoretical volume. It was written between 1899 and 1903, and then submitted as a dissertation to the University of Munich. Bouniatian's business cycle theory attracted some attention at the time (see, for example, Mitchell 1913, pp. 9–10; Keynes 1930, pp. 143–4) and his books were reviewed in the *Journal of the Royal Statistical Society* (June and September 1908), *American Economic Review* (June 1927, June 1936, December 1959), *Economic Journal* (September 1927, September 1932) and *Journal of Political Economy* (October 1934), among others.

Bouniatian's mix of theory and history in his *Studien* followed the pattern set by Mikhail Tugan-Baranovsky in his influential book about economic crises in England, published in Russian in 1894 and in a revised version in German in 1901. However, Bouniatian rejected the main elements of Tugan-Baranovsky's theory, that is, the compatibility between capital accumulation and decreasing consumption in the long run, and the notion that, in the depression, unused savings take the form of a fund of 'free capital' that is invested later in the upward period. It was not difficult for Bouniatian to show that actual saving and investment can never differ, although he did not consistently distinguish between desired and actual saving and investment – nor did Tugan–Baranovsky and most other contemporary economists for that matter. Concerning the first point, Bouniatian, building on Lauderdale (1804) and Mummery and Hobson (1889), carefully developed the view that there is in equilibrium a

certain relation between aggregate consumption and the capital stock determined by the choice of production methods, which he called 'degree of social capitalization' ('Grad der gesellschaftlichen Kapitalisierung'). This comes from Bouniatian's argument – against both Tugan-Baranovsky and the classical economists – that productive forces cannot be transferred to the future through the simple accumulation of capital goods, since these can be economically conserved only by being utilized in the process of production and sale of consumption goods.

According to Bouniatian, the evolution of the demand for investment through time is governed primarily by the evolution of consumers' 'new requirements' as determined by population growth, changes in tastes and inventions. However, this cannot be a smooth process because of the characteristics of the saving function on one side and of the production process of capital goods on the other. From the savers' side, whenever income grows there is a tendency – suggested by economic theory and confirmed by data – to increase the proportion of income saved. This 'tendency toward excessive accumulation' means that the demand for consumption goods tends to increase more slowly than production capacity, since saving is a 'false demand'. Such a tendency is realized due to the existence of a period of time necessary for the production of capital goods, which allows for a temporary separation between investment and consumption decisions and a more than proportional increase of capital goods in relation to a given intensification of 'new requirements', until the processes of production mature and consumers' good start to pour out. This was an early formulation of an aspect of what would later become known as the acceleration principle. 'Overcapitalization' ('Ueberkapitalisation', a term apparently coined by Bouniatian) is the main feature of the boom, which is followed by 'decapitalization' in the depression period, when overproduction of consumers' goods brings about a more than proportional fall in the value of capital goods. Equilibrium between production and consumption is restored through falling prices and depreciation of stocks and industrial plant, which transfer part of the capital to the consumption flow. However, equilibrium will not be attained if money wages are rigid downwards, as claimed by Bouniatian in his interpretation of the Great Depression of the early 1930s.

Apart from the saving function and the accelerator, another important element of Bouniatian's framework is his attempted application of the subjective theory of value to explain price level changes and, by that, the possibility of general overproduction. This was developed in detail in his 1927 book, where he used the Weber–Fechner law to generalize the old King's law – that the price of an important good varies inversely in geometrical progression as its quantity varies in arithmetical progression – to the economy as a whole. Bouniatian argued that, instead of the traditional quantity theory of money, price fluctuations should be explained by changes in the 'absolute social value' (marginal utility) of both consumption and capital goods, brought about by changes in their quantities throughout the business cycle. Such price changes are accompanied by changes in income distribution and, therefore, in the saving flow. This was used by Bouniatian (1908, vol. 1) to distinguish, for the first time in the literature, between 'exogenous' and 'endogenous' theories of the business cycle. In the latter, economic crises are explained as an organic part (the upper turning point) of the business cycle, not as accidents of economic history.

See Also

- ▶ Acceleration Principle
- ▶ Tugan-Baranovsky, Mikhail Ivanovich (1865–1919)

Selected Works

1908. *Studien zur Theorie und Geschichte der Wirtschaftskrisen,* 2 vols. Vol. 1: *Wirtschaftskrisen und Ueberkapitalisation – Eine Untersuchung über die Erscheinungsformen und Ursachen der periodischen Wirtschaftskrisen.* Vol. 2: *Geschichte der Handelskrisen in England im*

Zusammenhang mit der Entwicklung des englischen Wirtschaftslebens 1640–1840. Munich: E. Reinhardt.
1915. *Economic Crises – Morphology and Theory of Periodic Economic Crises and Theory of the Economic Conjuncture* [in Russian]. Moscow: N.P. Mesnyankin and Co.
1922. *Les crises économiques – essai de morphologie et théorie des crises économiques périodiques et de la conjoncture économique.* Trans. Russian by J. Bernard and revised by the author. Paris: M. Giard.
1927. *La loi de variation de la valeur et les mouvements généraux des prix.* Paris: M. Giard.
1928. The theory of economic cycles based on the tendency towards excessive capitalization. *Review of Economic Statistics* 10: 67–79. Reprinted in *Business cycle theory – Selected texts 1860–1939,* ed. M. Boianovsky, vol. 6. London: Pickering and Chatto, 2005.
1930. *Les crises économiques—essai de morphologie et théorie des crises économiques périodiques et de la conjoncture économique.* 2nd edn. Paris: M. Giard.
1933. *Crédit et conjoncture.* Paris: M. Giard.
1934. Economic depression and its causes. *International Labour Review* 30: 1–22.
1959. *Les fluctuations économiques – recueil d'études.* Paris: Pichon et Durand-Auzias.
1966. *Mes théories économiques et Albert Aftalion.* Paris: Pichon et Durand-Auzias.

Bibliography

Aftalion, A. 1908–9. La réalite des surproductions générales: essai d'une théorie des crises générales périodiques. *Revue d'Économie Politique* 22: 696–706; 23: 81–117, 201–29, 241–59. Trans. R. Leverdier in *Business cycle theory – Selected texts 1860–1939,* ed. M. Boianovsky, vol. 6. London: Pickering and Chatto, 2005.
Keynes, J.M. 1930. *A treatise on money*, vol. 1. London: Macmillan.
Lauderdale, J. 1804. *An inquiry into the nature and origin of public wealth, and into the means and causes of its increases*. Edinburgh: Archibald Constable.
Mitchell, W. 1913. *Business cycles*. Berkeley: University of California Press.
Mummery, A., and J. Hobson. 1889. *The physiology of industry*. London: J. Murray.
Tugan-Baranovsky, M. 1894. *Studien zur Theorie und Geschichte der Handelskrisen in England*. Jena: Fischer, 1901.

Bourgeoisie

J. Foster

The term *bourgeoisie* originally referred to the legal status of the town citizen in feudal France. In the *Encyclopédie* Diderot contrasted the political subordination of the *citoyen bourgeois* with the self-governing *citoyen magistrat* of ancient Greece. At the same time the French *bourgeoisie* (this term was first used in the 13th century) possessed certain economic and social rights, implicitly associated with the property required for trade, that distinguished it from the ordinary urban inhabitant or *domicilié* (Diderot 1753, III, 486–9).

Something of the same concept can be found in Hegel's use of the term *bürgerliche Gesellschaft* ('civil society'). Civil society represented the legal and governmental framework required for the 'actual achievement of selfish ends', the independent sphere of activity for the economic individual. It was in contrast to what Hegel saw as the embodiment of 'absolute rationality', the State, representing the universal interest of the whole community (Hegel 1820, p. 247).

Marx inherited, and initially used, *bourgeois* and *bürgerlich* in this restricted sense. Writing in 1842 on the opposition of the Rhineland urban estates to press freedom, he commented: 'we are faced here with the opposition of the bourgeois, not of the citoyen' (Marx 1842, p. 168). The petty and philistine motivation of the bourgeois is contrasted with the revolutionary impulses of the wider *Tiers Etat* as defined, for instance, by Siéyes (1789). By 1843–4, however, Marx had adopted an analysis of social change in terms of economically defined class forces and consequently identified the bourgeoisie, rather than an

undifferentiated *Tiers Etat*, as the revolutionary force which transformed feudal France. 'The negative general significance of the French nobility and the French clergy defined the positive general position of the immediately adjacent and opposed class of the *bourgeoisie'* (Marx 1844, p. 185). Four years later Marx gave classic expression to this historically progressive role in the *Communist Manifesto:*

> The bourgeoisie, during its rule of scarce one hundred years, has created more massive and more colossal productive forces than all preceding generations together . . . what earlier century had even a presentiment that such productive forces slumbered in the lap of social labour? (Marx 1848, p. 489).

At the same time, Marx also made a historically specific redefinition of *bürgerlich* or civil society. Civil rights, far from being abstract freedoms which derived from the political character of the State, in fact expressed the material interests of a class, the private owners of capital, and it was these that ultimately determined the nature of the State. 'The political revolution against feudalism' regarded the sphere of civil society as 'the basis of its existence'. Man 'was not freed from property, he received the freedom to own property' (Marx 1844, p. 167).

The crux of Marx's innovation was, therefore, to reconceive the terms bourgeoisie and bourgeois society in forms which anchored them to a particular mode of production. In the *Manifesto* the bourgeoisie is used as a synonym for capital ('the bourgeoisie, i.e. capital') while the 'executive of the modern state' is described as 'but a committee for managing the common affairs of the bourgeoisie as a whole' (Marx 1848, pp. 63 and 69).

Within this usage Marx invariably presents the bourgeoisie as historically contingent and subject to 'the immanent laws of capitalist production': to the 'centralisation of capital' and the contradictions bound up in its social relationship to labour. 'One capitalist kills many. Hand in hand with this centralisation, of the expropriation of many capitalists by few, develop on an ever extending scale, the co-operative form of the labour process . . .' (Marx 1867, p. 714–15). Accordingly, as Marx stressed in his *Eighteenth Brumaire of Louis Napoleon,* an analysis of the bourgeoisie, and of its internal 'factions' and 'interests', had to start with a concrete assessment of its particular forms of property and their changing place within capitalist production: 'upon the different forms of property, upon its social conditions of existence, rises an entire superstructure of distinct and differently formed sentiments . . .' (Marx 1852, p. 128).

The petty bourgeoisie, for instance, represented an unstable and transitional layer between the bourgeoisie and the proletariat:

> in countries where modern civilisation has become fully developed, a new class of petty bourgeoisie has been formed, fluctuating between proletariat and bourgeoisie and ever renewing itself as a supplementary part of bourgeois society . . . as modern industry develops, they even see the moment approaching when they will completely disappear as an independent section of modern society and be replaced . . . by overseers, bailiffs and shop assistants (Marx 1848, p. 509).

They represented a 'transitional class in which the interests of two classes are simultaneously mutually blunted . . .' (Marx 1852, p. 133).

Conversely, within the bourgeoisie the centralization of capital ultimately reaches a point where management and ownership become divorced: 'the transformation of the actually functioning capitalist into a mere manager, an administrator of other people's capital and of the owner of capital into a mere owner, a mere money capitalist . . ..'

Credit offers to the individual capitalist . . . absolute control over the capital and property of others . . . and thus to expropriation on the most enormous scale. Expropriation extends here from the direct producers to the smaller and medium-sized capitalists themselves . . ..

But 'instead of overcoming the antithesis between the character of wealth as social or a private wealth, the stock companies merely develop it in a new form' (Marx [1894], 1959, pp. 436–41).

Hence, in sum, Marx radically extended the significance of the concept to make the bourgeoisie that class which produced, but was itself continually modified by, the capitalist mode of production. Conversely, Marx gave a new and historically specific meaning to the term 'civil' (or *bürgerlich*) society, and argued that its endorsement of individual liberties extended

only so far as they were compatible with capitalist property relations.

In the following generation a number of notable non-Marxist scholars adopted, at least in part, Marx's identification of the bourgeoisie as the class responsible for winning the social and political conditions necessary for capitalist production. But this process of wider adoption also saw a further reorientation of the concept. The new political and social institutions created by the bourgeoisie were now presented as the definitive basis for human freedom. The bourgeois character of civil society became the ultimate justification for the bourgeoisie.

Pirenne, writing in the 1890s, traced back the personal liberties of modern society to the medieval merchant bourgeoisie. It was the reliance of this class of merchant adventurers on individual enterprise and the unfettered application of knowledge that made the bourgeoisie the universal champion of 'the idea of liberty' (Pirenne 1895, 1925).

A little later Weber identified the origins of capitalist enterprise in the rational, resource-maximizing practices of medieval book-keeping. He then went one step further to claim that this 'capitalist spirit' was in turn derived from the doctrines of individual responsibility and conscientious trusteeship found in early protestant theology. Parallel to this within the political sphere, Weber argued that the same doctrines also underlay the creation of representative institutions and constitutional government (Weber 1901–2, 1920).

In the 1940s Schumpeter extended this derivation to democracy itself: 'modern democracy is a product of the capitalist process' (Schumpeter 1943, p. 297). To do so he redefined the essence of democracy in individual, market terms as 'free competition for a free vote' (1943, p. 271), and warned that this was likely to be destroyed unless the advance of socialism could be halted. Schumpeter's thesis has since been generalized by Barrington Moore, who has sought to demonstrate that all forms of social modernization *not* led by the bourgeoisie have produced totalitarian forms of government (Moore 1969).

This redefinition of Marx's original usage is also found in the continuing debate on the transition from feudalism to capitalism. Paul Sweezy, following Pirenne, argued that it was trade, and the role of the urban bourgeoisie as merchants, that destroyed feudalism as a mode of production. Towns and trade were alien elements that had corroded feudalism's non-market, nonexchange modes of appropriation (Sweezy 1950). Maurice Dobb, following Marx's usage, had previously sought to show that the medieval bourgeoisie only became a revolutionary class in so far as it challenged feudalism as a mode of production (not distribution) and attempted to create a new type of exploitative relationship between capital and proletarianized labour (Dobb 1946, p. 123, 1950). Dobb referred to Marx's own contention that the fully revolutionary overthrow of feudalism only took place when the struggle was under the leadership of the 'direct producers' rather than the merchant elite (Marx [1894], 1959, pp. 327–37).

Recently Anderson has revived this argument in a new form. Seeking the origins of the non-absolutist and democratic forms of government found in Western Europe, he argued that such institutions depended on a 'balanced fusion' between the feudalized rural remnants of Germanic society and the urban heritage of Roman *civiltas* and contract law. The role of the medieval merchant bourgeoisie within this fusion was to act as the bearer of the urban tradition (Anderson 1974; see also Brenner 1985).

The other major area of redefinition has been directed at the bourgeoisie in late or 'post' capitalist society. Its central feature is the claimed separation between the ownership and management of capital. If the bourgeoisie is defined by an ownership of capital that involves effective possession and control (Balibar 1970), it is argued that in modern industrial society the actual owners of capital, the shareholders, have surrendered this to a 'new class' of corporate managers (Gouldner 1979; Szelenyi 1985). This concept of a managerial revolution was first popularized by Burnham (1942). It has since been developed to take account of the transnational concentration of capital. The resulting specialization of company functions has, it is argued, given executives the power to create autonomous spheres of decision-making with the result that corporate goals and strategies

do not necessarily reflect the profit-maximizing interests of the nominal owners (Chandler 1962; Pahl and Winkler 1974).

In contrast, Marx has contended in his final writings that the growth of industrial monopoly and credit heightened the contradiction between private ownership and social labour, distorted exchange relationships and demanded systematic state intervention (Marx [1894], 1959, p. 438). Lenin later elaborated this perspective to argue that the growth of monopoly marked a new and final stage of capitalist development in which a fundamental split took place within the bourgeoisie. Utilizing an analysis first made by Hilferding (1910), Lenin argued that the fusion of banking and monopoly capital, producing 'finance capital', had created a new and parasitic relationship between state power and just one section of the bourgeoisie. The result was 'state monopoly capitalism' (Lenin 1916, 1917). A recent variant of this analysis has used the interlocking of company directorships to argue for the existence of a controlling elite of directors exercising a strategic dominance over all capital (Aaronovitch 1961; Useem 1984; Scott 1984).

See Also

- ▶ Capitalism
- ▶ Class

References

Aaronovitch, S. 1961. *The ruling class*. London: Lawrence & Wishart.

Anderson, P. 1974. *Lineages of the Absolutist State*. London: New Left Books.

Balibar, E. 1970. Basic concepts of historical materialism. In *Reading capital*, ed. L. Althusser and E. Balibar. London: New Left Books.

Brenner, R. 1985. Agrarian class structure and economic development in pre-industrial Europe. In *The Brenner debate*, ed. T. Aston. Cambridge: Cambridge University Press.

Burnham, J. 1942. *The managerial revolution*. London: Putnam.

Chandler, A. 1962. *Strategy and structure*. Cambridge, MA: MIT Press.

Diderot, D. 1753. *Encyclopédie ou Dictionnaire raisonné des sciences*. Paris: Briasson.

Dobb, M. 1946. *Studies in the development of capitalism*. London: Routledge.

Dobb, M. 1950. A reply. *Science and Society*.

Gouldner, A. 1979. *The future of intellectuals and the rise of the new class*. New York: Seabury Press.

Hegel, G. 1820. Naturrecht und Staatswischenschaft in Grundrisse. In *Werke*, vol. VIII, ed. E. Gans. Berlin.

Hilferding, R. 1910. *Das Finanzkapital*. Vienna: I. Brand. Trans. M. Watnick and S. Gordon as *Finance capital,* ed. T. Bottomore. London: Routledge & Kegan Paul, 1981.

Lenin, V.I. 1916. Imperialism: The highest stage of capitalism. In *Collected works*, vol. XXIII. Moscow: Progress, 1964.

Lenin, V.I. 1917. The impending catastrophe and how to combat it. In *Collected works*, vol. XXV. Moscow: Progress, 1964.

Marx, K. 1842. Debate on the law on thefts of wood. In *Collected works*, vol. I. Moscow: Progress, 1975.

Marx, K. 1844. Contribution to the critique of Hegel's Philosophy of Law. In *Collected works*, vol. III. Moscow: Progress, 1975.

Marx, K. 1848. The Manifesto of the Communist Party. In *Collected works*, vol. VI. Moscow: Progress, 1976.

Marx, K. 1852. The eighteenth Brumaire of Louis Napoleon. In *Collected works*, vol. XI. Moscow: Progress, 1976.

Marx, K. 1867. *Capital*, vol. I. Moscow: Progress, 1953.

Marx, K. 1894. *Capital*, vol. III. Moscow: Progress, 1959.

Moore, B. 1969. *The social origins of democracy and dictatorship*. London: Penguin.

Pahl, R., and J. Winkler. 1974. The economic elite: Theory and practice. In *Elites and power in British society*, ed. P. Stanworth and A. Giddens. Cambridge: Cambridge University Press.

Pirenne, H. 1895. L'origine des constitutions urbaines au moyen age. *Revue historique* 57.

Pirenne, H. 1925. *Medieval Cities: Their origin and the renewal of trade*. Princeton: Princeton University Press.

Schumpeter, J. 1943. *Capitalism, socialism and democracy*. London: George Allen & Unwin.

Scott, J. 1984. *Directors of industry: The British corporate network*. Cambridge: Polity Press.

Siéyes, E. 1789. *Qu'est-ce que le Tiers Etat?* Paris.

Sweezy, P. 1950. The transition from feudalism to capitalism. *Science and Society* 14(2): 134–157.

Szelenyi, I. 1985. Social policy and state socialism. In *Stagnation and renewal in social policy*, ed. G. Esping-Anderson. White Plains: Sharpe.

Useem, M. 1984. *The inner circle*. New York: Oxford University Press.

Weber, M. 1901–2. Die protestantische Ethik und der Geist des Kapitalismus. *Archiv für Sozialwissenschaft und Sozialpolitik* 20.

Weber, M. 1920. *Gessämmelte Aufsätze zur Religionssoziologie*. Tübingen: Mohr.

Bowley, Arthur Lyon (1869–1957)

J. R. N. Stone

Keywords

Bowley, A. L.; Econometric Society; Family budgets; Fisher, I.; Ideal index numbers; Index numbers; International Statistical Institute; Marshall, A.; Mathematical economics; *Palgrave's Dictionary of Political Economy*; Royal Statistical Society; Sampling; Statistics and economics

JEL Classifications

B31

Bowley was born on 6 November 1869 in Bristol, and died on 21 January 1957 at Haslemere. In 1922 he was made a Fellow of the British Academy and knighted in 1950. He was educated at Christ's Hospital from 1879 to 1888, and Trinity College, Cambridge, from 1888 to 1891 (10th Wrangler, 1891). He stayed on another two terms studying physics, chemistry and, under the influence of Alfred Marshall, who remained a lifelong friend, economics. After a period as a schoolmaster, he became lecturer in mathematics, and then professor of mathematics and economics at University College, Reading, from 1900 to 1919. He concurrently taught at the London School of Economics from its inception in 1895, first as lecturer, then reader, then professor, and finally, from 1919, as the first holder of the newly established Chair of Statistics at the University of London, becoming Emeritus Professor on his retirement in 1936.

Among his other activities, he was Acting Director of the Oxford University Institute of Statistics from 1940 to 1944; foundation member in 1933, and then President from 1938 to 1939, of the Econometric Society; President of the Royal Statistical Society from 1938 to 1940, and honorary President of the International Statistical Institute in 1949.

Bowley was an outstanding economic statistician who made substantial contributions to all areas in his field, from the theory of mathematical statistics to the methodology and practice of data collecting. His courses on statistics at the LSE formed the subject matter of two very successful textbooks (Bowley 1901, 1910). He brought together and set out in a uniform way the developments of mathematical economics from Cournot to Pigou (Bowley 1924). He wrote a detailed account of Edgeworth's contributions to mathematical statistics (Bowley 1928). He collaborated with R.G.D. Allen on a masterly study of family budgets which deals with individual variation as well as average behaviour (Allen and Bowley 1935).

One of his early interests was the course of wages, on which he wrote several books and over 30 articles, many jointly with G.H. Wood; his first paper on the subject was Bowley (1895) and his first book Bowley (1900). This led him to write extensively on index-numbers of prices and it is interesting that in 1899, on p. 641 of vol. III of *Palgrave's Dictionary of Political Economy*, he gave the index-number formula later to become famous as Irving Fisher's ideal index-number. He followed this work with studies of the national income in Bowley (1919, 1920, 1937) and jointly with J.C. Stamp in Bowley and Stamp (1927).

Bowley was a pioneer in the development of sampling methods and spoke strongly in their favour in his presidential address to the British Association in 1906. In 1912 he carried out a well-designed sample survey of Reading and soon followed this with similar enquiries in Northampton, Warrington, Stanley and Bolton (Bowley and Burnett-Hurst 1915). A second survey of the same towns was made after the war (Bowley and Hogg 1925). In the same period he prepared a substantial report on the precision attained in sampling (Bowley 1926). He played an important role in Llewellyn-Smith's new survey of London life and labour (Bowley 1930–35).

Selected Works

1895. Changes in average wages (nominal and real) in the United Kingdom between 1860

and 1891. *Journal of the Royal Statistical Society* 58: 223–278.
1900. *Wages in the United Kingdom in the nineteenth century.* Cambridge: Cambridge University Press.
1901. *Elements of statistics.* London: P.S. King. 6th ed, 1937.
1910. *An elementary manual of statistics.* London: P.S. King. 7th ed, London: Macdonald & Evans, 1951.
1915. (With A.R. Burnett-Hurst.) *Livelihood and poverty: A study in the economic conditions of working-class households in Northampton, Warrington, Stanley and Reading*. London: G. Bell.
1919. *The division of the product of industry: An analysis of national income before the war.* Oxford: Clarendon Press.
1920. *The change in the distribution of the national income, 1880–1913.* Oxford: Clarendon Press.
1924. *The mathematical groundwork of economics*. Oxford: Clarendon Press.
1925. (With M.H. Hogg.) *Has poverty diminished?* London: P.S. King.
1926. Measurement of the precision attained in sampling. *Bulletin de l'Institut International de Statistique* 22, pt I(3): 1–62.
1927. (With J.C. Stamp.) *The national income 1924.* Oxford: Clarendon Press.
1928. *F.Y. Edgeworth's contributions to mathematical statistics.* London: Royal Statistical Society.
1930–35. Contributions to H. Llewellyn-Smith. In *New survey of London life and labour*, 9 vols. London: P.S. King.
1935. (With R.G.D. Allen.) *Family expenditure.* London: P.S. King.
1937. *Wages and income in the United Kingdom since 1860.* Cambridge: Cambridge University Press.

Bibliography

Allen, R.G.D. 1968. Bowley, Arthur Lyon. In *International encyclopedia of the social sciences*, vol. 2. New York: Macmillan and Free Press.
Allen, R.G.D. 1971. Bowley, Sir Arthur Lyon. In *Dictionary of national biography, 1951–1960*. Oxford: Oxford University Press.
Allen, R.G.D., and George, R.F. 1957. Professor Sir Arthur Lyon Bowley (with bibliography). *Journal of the Royal Statistical Society*, Series A (General) 120(pt 2): 236–241.
Bowley, Agatha H. 1972. *A memoir of Professor Sir Arthur Bowley* (1869–1957) *and his family.* Privately printed.
Darnell, A. 1981. A.L. Bowley, 1869–1957. In *Pioneers of modern economics in Britain*, ed. D.P. O'Brien and J.R. Presley. London: Macmillan.
Maunder, W.F. 1977. Sir Arthur Lyon Bowley. In *Studies in the history of statistics and probability*, vol. 2, ed. M.G. Kendall and R.L. Plackett. London: Griffin.

Bowley, Marian (Born 1911)

B. A. Corry

Abstract

Marian Bowley was born in 1911, the daughter of the distinguished statistician A.L. Bowley. She was a student at the London School of Economics (1928–31), where she took her BSc (Econ) degree and later her PhD in 1936. She held a series of temporary teaching and research posts and was appointed to a lectureship at the Dundee School of Economics in 1938. After government service during World War II she was appointed to a lectureship at University College, London in 1947 and became successively reader and professor. She retired in 1975 and was made professor emeritus.

Marian Bowley was born in 1911, the daughter of the distinguished statistician A.L. Bowley. She was a student at the London School of Economics (1928–31), where she took her BSc (Econ) degree and later her PhD in 1936. She held a series of temporary teaching and research posts and was appointed to a lectureship at the Dundee School of Economics in 1938. After government service during World War II she was appointed to a lectureship at University College, London in 1947 and became successively reader and professor. She retired in 1975 and was made professor emeritus.

Marian Bowley's best-known contribution to economics is her work in the history of economic

thought. Her major work in this field is undoubtedly her *Nassau Senior and Classical Economics* (1937) which still remains the standard work on that much misunderstood member of the classical school. Her book is more than just a study of Senior, it is really an overview of the whole classical system, both its economic theory and policy stance, woven into a study of Senior. One of her major points was to question the hegemony of classical value theory and argue rather that there were two distinct strands: the labour theory propagated by the Ricardians and a subjective approach espoused by people such as Lauderdale and Senior.

Bowley's other contributions to the history of economics are collected in her *Studies in the History of Economic Theory before 1870* (1973), where, incidentally she somewhat repudiates her earlier views on classical value theory and sees more common features in the analysis.

Marian Bowley has also made important contributions to the understanding of the building industries in her *Innovations in Building Materials* (1960) and *The British Building Industry* (1966).

Selected Works

1937. *Nassau senior and classical economics*. London: George Allen & Unwin.
1960. *Innovations in building materials*. London: Duckworth.
1966. *The British building industry*. Cambridge: Cambridge University Press.
1973. *Studies in the history of economic theory before 1870*. London: Macmillan.

Bowman, Mary Jean (Born 1908)

Gilbert R. Ghez

Mary Jean Bowman, born in 1908 in New York City, obtained her Ph.D. at Harvard (1938). Since 1958 she has taught at the University of Chicago. Her publications include ten books and monographs and over 100 articles, most of which relate to the economics of education.

Bowman's early writings are primarily expository: a clear description of measures of income inequality, and a stimulating textbook on economics written jointly with G.L. Bach. Much of her later writing deals with the effects of education on economic development and the personal distribution of income, using US, Japanese, Malaysian and Mexican data. These studies, which relate to schooling and on-the-job training, are well documented. Bowman also emphasizes the role of fertility and technological change, arguing that high rates of human capital formation and high rates of population growth are incompatible, unless sustained by technological change and the ability to learn rapidly in the post-school years.

Two characteristics typify her writings. First is her repeated attempts to bring expectations and uncertainty to bear on educational choices, using concepts inspired by G.L.S. Shackle. Secondly, Bowman often uses a multi-disciplinary approach. A contributor to the human capital and home economics literature, she draws on the education and sociological literature as well. An example of her tribute to sociological ideas is her insistence on the importance of 'information fields'. It remains to be seen whether these often disparate concepts, which Bowman has juggled so successfully, can be formalized in a synthetic fashion.

Selected Works

1943. (With G.L. Bach). *Economic analysis and public policy*. New York: Prentice-Hall.
1945. A graphical analysis of personal income distribution in the United States. *American Economic Review* 35(3): 607–628. September.
1958. (ed.) *Expectations, uncertainty and business behavior*. New York: Social Science Research Council.
1963. (With W.W. Haynes). *Resources and people in East Kentucky: Problems and prospects of a*

lagging economy. Baltimore: Johns Hopkins Press for Resources for the Future.
1964. Schultz, Denison, and the contribution of 'Eds' to national income growth. *Journal of Political Economy* 72(5): 450–464. October.
1965. (ed., with C.A. Anderson). *Education and economic development*. Chicago: Aldine.
1966. The costing of human resource development. In *The economics of education*, ed. E.A.G. Robinson, J. Vaizey. London: Macmillan.
1967. (With R.G. Myers). Schooling, experience, and gains and losses in human capital through migration. *Journal of the American Statistical Association* 62(3): 875–898. September.
1968. (ed.) *Readings in the economics of education*. Paris: UNESCO.
1972. (ed., with C.A. Anderson and V. Tinto). *Where colleges are and who attends*. New York: McGraw Hill.
1972. Expectations, uncertainty and investments in human beings. In *Uncertainty and expectations in economics*, ed. C.F. Carter, J.L. Ford. Oxford: Basil Blackwell.
1973. (With D. Plunkett). *Elites and change in the Kentucky mountains*. Lexington: University of Kentucky Press.
1978. (With A. Sohlman and B-C. Ysander). *Learning and earning*. Stockholm: National Board of Universities and Colleges.
1981. (With the collaboration of H. Ikeda and Y. Tomoda). *Educational choice and labor markets in Japan*. Chicago: University of Chicago Press.
1982. Choice in the spending of time. In *The social sciences, their nature and uses*, ed. W.H. Kruskal. Chicago: University of Chicago Press.
1984. An integrated framework for analysis of the spread of schooling in less developed countries. *Comparative Education Review* 28(4): 563–583. November.
1986. (With B. Millot and E. Schiefelbein). *Political economy of public support of higher education: Studies in Chile, France and Malaysia*. Washington, DC: World Bank. Discussion Paper.

Boyd, Walter (1754–1837)

David Laidler

B

Keywords
Bank of England; Boyd, W.; Bullionist controversy; Convertibility; Sinking Fund

JEL Classifications
B31

Before the French Revolution, Walter Boyd was engaged as a banker in France, but by the time his firm's property was confiscated by the French government in 1793, he was established in London as the leading member of the firm of Boyd Benfield & Co. At first this London venture was highly successful, and in 1797 Boyd entered Parliament as member for Shaftsbury, then a pocket borough owned by his partner. In this very year, however, Boyd Benfield & Co began to encounter the difficulties which were to culminate in its liquidation in 1800. The basic cause of Boyd's ruin was his having entered into engagements in the expectation that his French property would be restored to him, an expectation that was finally disappointed in September 1797, but the events which precipitated the final collapse of his firm were the government's refusal to employ it as a contractor for the loan of 1799 and the Bank of England's final refusal to grant assistance in early 1800.

When, in 1801, Boyd published his 'Letter to William Pitt. . .' attacking the Bank of England's policies since the suspension of specie convertibility of February 1797, he was hardly a disinterested observer. However, this pamphlet's appearance is widely regarded as marking the beginning of the 'Bullionist Controversy', and contains perhaps the first systematic, albeit crude, statement of what came to be known as the Bullionist position. It argued that exchange depreciation and food price increases since 1797

were the result of an overissue of paper money by the Bank of England; that though foreign transfers could depreciate the exchanges this factor had not been important since 1797; and that the Country Bank note issue could not affect prices independently of Bank of England policies.

Boyd's pamphlet drew a number of replies, some, as Fetter (1965) notes, aimed more at Boyd than at his case, but one by Sir Francis Baring (1801) prefigured subsequent anti-bullionist positions. Baring argued (with some justice) that food price behaviour had had more to do with bad harvests than the exchange rate (which had moved much less), and that the exchange rate's fall had been the result of British remittances to Continental allies and not of over-expansionary policy on the part of the Bank of England.

Boyd made no further contributions to wartime debates. After the Peace of Amiens (1802) he visited France, only to be trapped there until 1814 by the renewal of hostilities. Upon his return to England he re-established his fortunes sufficiently to be able to re-enter Parliament in 1823, as member of Lymington, which he represented until 1830. He published two further pamphlets, on the Sinking Fund (1815 and 1828), but neither of these has the historical significance of his 1801 contribution.

See Also

► Bullionist Controversies (Empirical Evidence)

Selected Works

1801. *Letter to the right honourable William Pitt on the influence of the stoppage of issues in specie at the Bank of England on the Prices of provisions and other commodities*. London. 2nd ed., 1811.
1815. *Reflections on the financial system of great britain, and particularly on the sinking Fund.* London. 2nd ed., 1828.
1828. *Observations on Lord Grevilles essay on the sinking Fund.* London.

Bibliography

Baring, Sir F. 1801. *Observations on the publication of Walter Boyd.* London.
Fetter, F.W. 1965. *Development of British monetary orthodoxy: 1797–1875*. Cambridge, MA: Harvard University Press.

Bradford, David (1939–2005)

Daniel N. Shaviro

Keywords

Bradford, D.; Consumption taxation; Flat tax; Intertemporal preferences; Local public finance; Progressive and regressive taxation; Public goods pricing; Taxation of corporate profits; Taxation of income

JEL Classifications

B31; H21; H22; H24; H25; H73

David Bradford is best known for his work on fundamental tax reform, although his contributions to public economics were more wide-ranging. His early writings, after he joined the economics department at Princeton University in 1966, largely focused on municipal finance and public goods pricing. His interests took a dramatic turn, however, when he was named Deputy Assistant Secretary for Tax Policy at the US Treasury in 1975, and given lead responsibility for a Treasury study of comprehensive tax reform. The influential *Blueprints for Basic Tax Reform* (1977), which he co-authored with the US Treasury Tax Policy Staff, set forth models for comprehensive income and consumption taxes that remain influential to this day. The *Blueprints* cash flow consumption tax in particular influenced subsequent tax reform thinking by showing how a consumption tax, levied at the individual rather than the business level, could match the progressivity of an income tax and offer self-help income averaging

through a mix of 'prepaid' and 'postpaid' (that is, yield-exempt and deductible) savings accounts.

His experiences at the Treasury made Bradford a lifelong advocate of consumption taxation, based on two main considerations. The first was that he considered it inequitable for people with the same lifetime earnings to face different tax burdens, as they would under an income tax, simply by reason of having different intertemporal consumption preferences. The second was that shifting to a consumption base might permit significant tax simplification, by eliminating the timing issues that bedevil a realization-based income tax. Bradford later developed a second consumption tax prototype, the X-tax, based on the Hall–Rabushka flat tax (Hall and Rabushka, 1995) but modified to permit greater progressivity and to address transition problems, which he recognized could arise not only upon initial enactment but whenever tax rates were changed.

Bradford also helped to pioneer the contemporary understanding that the only theoretical difference between pure income and consumption taxation lies in their treatment of the risk-free return to waiting, which the former subjects to tax and the latter exempts. In addition, he advanced understanding of the economics of a transition from income to consumption taxation, showing that the ostensibly lump-sum revenue gain resulted from wiping out assets' income tax basis while solemnly pledging never to do so again. Bradford also helped develop the 'new view' of corporate taxation, which shows that a uniform tax on corporate distributions does not distort corporate decisions regarding when to pay out earnings.

See Also

- ▶ Consumption taxation
- ▶ Taxation of corporate profits
- ▶ Taxation of income

Selected Works

1970. (With W. Baumol.) Optimal departures from marginal cost pricing. *American Economic Review* 60: 265–283.

1977. (With US Treasury Tax Policy Staff.) *Blueprints for basic tax reform*. Arlington, VA: Tax Analysts, 1984.

1981. The incidence and allocation effects of a tax on corporate distributions. *Journal of Public Economics* 15: 1–22.

1986. *Untangling the income tax*. Cambridge, MA: Harvard University Press.

1996. Consumption taxes: Some fundamental transition issues. In *Frontiers of tax reform*, ed. M. Boskin. Stanford, CA: Hoover Institution Press.

2004. (With A. Auerbach.) Generalized cash-flow taxation. *Journal of Public Economics* 88: 957–980.

Bibliography

Hall, R., and A. Rabushka. 1995. *The flat tax*. Stanford, CA: Hoover Institution Press.

Brady, Dorothy Stahl (1903–1977)

M. Reid

Keywords

Brady, D. S.; Distribution of income; Survey data

JEL Classifications

B31

A mathematician and statistician, Dorothy S. Brady combined in her professional life extended periods in both universities and US federal agencies. Most of her empirical work entailed the design and interpretation of survey data on household income and expenditures and critiques of applications of such data.

This began with analysis of data collected in the large 1935–6 survey of incomes and expenditures of rural households which together with its urban counterpart provided the basis for new tests of the validity of Commerce Department

estimates of the size and distribution of national income, consumption and savings. At the Bureau of Labor Statistics (1943–8, 1951–6) she assessed consumption and price data in connection with efforts to control inflation, and she developed the statistical design for pricing the city workers' family budget which was used to estimate inter-area differences in the cost of living.

An active participant in the Conference on Income and Wealth of the National Bureau of Economic Research, Brady brought to its sessions a keen awareness of data limitations in the empirical identification of key elements in an analytical structure. Using statistical analysis to randomize effective unidentified factors, she found that the percentage of income saved by families tends to increase systematically with relative position in an income distribution, that the secular increase of income of a population tends to decrease the age at which children leave the family residence, often with financial help from parents, and that such leaving tends to increase the inequality of measured income distribution.

Selected Works

1940. (With others). *Family income and expenditures, 5 regions,* Part 2. Miscellaneous publication no. 396. Washington, DC: US Department of Agriculture.

1945. *Family spending and saving in wartime.* US Bureau of Labour Statistics, Bulletin No. 872, Appraisal of survey data, p. 45.

1948. (With L.S. Kellog). The city worker's family budget. *Monthly Labour Review* 66(25): 133–70.

1949. The use of statistical procedures in the derivation of family budgets. *Social Service Review* 23(2): 141–57.

1952. Family savings in relation to changes in the level and distribution of income. In *Studies in income and wealth,* vol. 15. New York: NBER.

1956. Family savings, 1880 to 1950. Part II. In *A study of savings in the United States,* vol. III: *Special studies,* ed. R.W. Goldsmith, D.S. Brady and H. Mendershausen. Princeton: Princeton University Press.

1957. Measurement and interpretation of the income distribution in the United States. *Income and wealth*, series VI. London: International Association for Research in Income and Wealth.

1958. Individual incomes and the structure of consumer units. *American Economic Review, Papers and Proceedings* 48: 269–78.

1965. *Age and the income distribution.* Research report no. 8. Washington, DC: US Social Security Administration.

1966. Price deflators for final product estimates. In *Studies in income and wealth,* vol. 30: *Output, employment, and productivity in the United States after 1800.* Conference on Research in Income and Wealth. New York: NBER.

1971. The statistical approach: The input–output system. In *Approaches to American economic history,* ed. G.R. Taylor and L.F. Ellsworth. Charlottesville: University Press of Virginia.

Brain Drain

Frédéric Docquier and Hillel Rapoport

Abstract

The term 'brain drain' designates the international transfer of human resources and mainly applies to the migration of relatively highly educated individuals from developing to developed countries. While the brain drain has long been viewed as detrimental to poor countries' growth potential, recent economic research emphasizes a number of positive feedback effects arising from skilled migrants' participation in business networks, and suggests that under certain conditions the prospect of migration can positively affect human capital accumulation in the source countries.

Keywords

Altruism; Brain drain; Diaspora networks; Education; Externality; Foreign direct investment; Human capital; Information costs;

International migration; Remittances; Transaction costs; Transfer of technology; Wage differentials

JEL Classifications
F

The term 'brain drain' designates the international transfer of resources in the form of human capital and mainly applies to the migration of relatively highly educated individuals from developing to developed countries. In the non-academic literature, the term is generally used in a narrower sense and relates more specifically to the migration of engineers, physicians, scientists and other very highly skilled professionals with university training. The brain drain has long been viewed as a serious constraint on poor countries' development and is also a matter of concern for many European countries such as the UK, Germany or France, which have recently seen a significant fraction of their talented workforce emigrate abroad. Recent comparative data reveal that by 2000 there were 20 million highly skilled immigrants (that is, foreign-born workers with a tertiary education) living in the Organisation for Economic Co-operation and Development (OECD) area, a 70 per cent increase in ten years against only a 30 per cent increase for unskilled immigrants. Skilled migrants now represent one-third of total immigration to the OECD countries, and most of this increase is due to immigration from developing and transition countries. The causes of this growing brain drain are well known. On the supply side, the globalization of the world economy has strengthened the tendency for human capital to agglomerate where it is already abundant and has contributed to increase positive self-selection among migrants. And on the demand side, host countries have gradually introduced quality-selective immigration policies and are now engaged in what appears as an international competition to attract global talents.

How Big is the Brain Drain?

Extending and updating the work of Carrington and Detragiache (1998), Docquier and Marfouk (2006) recently collected OECD immigration data to construct estimates of emigration rates by educational attainment (primary, secondary and tertiary schooling) for all countries in 1990 and 2000. Their figures for the highest education level may be taken as a brain drain measure. This may seem too broad a definition for the most advanced countries where the highly educated typically represent about a third of the total workforce but seems appropriate in the case of developing countries, where this share is on average just about five per cent. Note that due to data constraints, South–South migration is not taken into account in the Docquier and Marfouk (2006) data-set; this can lead to potential underestimation of the brain drain for some countries for which other developing countries are significant destinations. On the other hand, the very definition of immigrants as foreign-born workers does not account for whether education has been acquired in the home or in the host country; this can lead to potential overestimation of the brain drain as well as to possible spurious cross-country variation in skilled emigration rates (Rosenzweig 2005). In an attempt to solve this problem, Beine et al. (2007a) used age of entry as a proxy for where education has been acquired and proposed alternative brain drain estimates excluding people who immigrated before a given age (12, 18 and 22); their results show country rankings by degree of brain drain intensity only mildly affected by the correction and extremely high correlations between corrected and uncorrected estimates.

Keeping this in mind, one can use a simple multiplicative decomposition of the brain drain: the skilled emigration rate is to equal to the average emigration rate times the schooling gap. The average emigration rate is the ratio of emigrants to natives (residents plus emigrants) and reflects the sending country's openness to emigration. The schooling gap is the ratio of skilled to average emigration rate which, by definition, is also the ratio of the proportion of educated among emigrants to the corresponding proportion among natives.

Table 1 summarizes the data for different country groups in 2000. Countries are grouped according to demographic size, income per capita

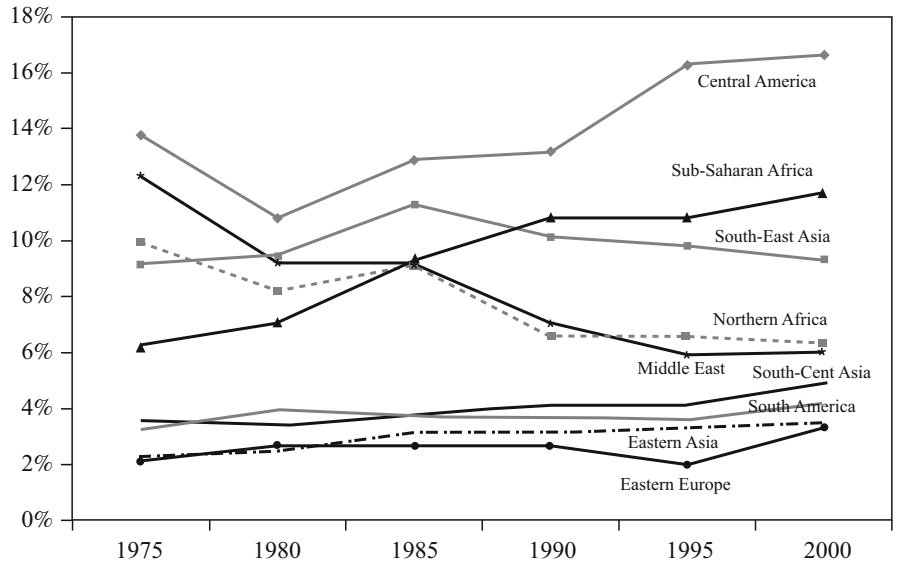
Brain Drain, Table 1 Data by country group in 2000

	Skilled emigration rate (%)	*Average emigration rate (%)*	*Schooling gap*
By population size (millions)			
Large countries (>25)	*4.1*	*1.3*	*3.144*
Upper-middle (>10–25)	*8.8*	*3.1*	*2.839*
Lower-middle (>2.5–10)	*13.5*	*5.8*	*2.338*
Small countries (< 2.5)	*27.5*	*10.3*	*2.666*
By income group			
High-income countries	*3.5*	*2.8*	*1.238*
Upper-middle income	*7.9*	*4.2*	*1.867*
countries			
Lower-middle income	*7.6*	*3.2*	*2.383*
countries			
Low-income countries	*6.1*	*0.5*	*12.120*
By region			
AMERICA	*3.3*	*3.3*	*1.002*
USA and Canada	*0.9*	*0.8*	*1.127*
Caribbean	*42.8*	*15.3*	*2.807*
Central America	*16.9*	*11.9*	*1.418*
South America	*5.1*	*1.6*	*3.219*
EUROPE	*7.0*	*4.1*	*1.717*
Eastern Europe	*4.3*	*2.2*	*1.930*
Rest of Europe	*8.6*	*5.2*	*1.637*
incl. EU15	*8.1*	*4.8*	*1.685*
AFRICA	*10.4*	*1.5*	*7.031*
Northern Africa	*7.3*	*2.9*	*2.489*
Sub-Saharan Africa	*13.1*	*1.0*	*13.287*
ASIA	*5.5*	*0.8*	*7.123*
Eastern Asia	*3.9*	*0.5*	*8.544*
South-central Asia	*5.3*	*0.5*	*10.030*
South-eastern Asia	*9.8*	*1.6*	*5.980*
Near and Middle East	*6.9*	*3.5*	*1.937*
OCEANIA	*6.8*	*4.3*	*1.578*
Australia and New Zealand	*5.4*	*3.7*	*1.479*
Other Pacific countries	*48.7*	*7.6*	*6.391*

Source: Docquier and Marfouk (2006)

(under the World Bank classification), and region. Unsurprisingly, we observe a decreasing relationship between emigration rates and country size, with average skilled emigration rates about seven times higher in small countries than in large countries. Regarding income groups, the highest emigration rates are observed in middle-income countries, where people have both the incentives and means to emigrate. Regarding the regional distribution of the brain drain, the most affected regions are the Caribbean and the Pacific islands, sub-Saharan Africa (where the schooling gap is exceptionally high), and Central America.

It is clear that the *magnitude* of the brain drain has increased dramatically since 1980. However, in terms of *intensity* (or emigration rates), the picture is less clear as one must factor in the general progress in educational attainments observed across the world. Figure 1 presents skilled emigration rates by region computed by



Brain Drain, Fig. 1 Long-run trends in skilled emigration, 1975–2000. *Source:* Defoort (2006)

Defoort (2006) using a long-run perspective. Focusing on the six major destination countries (USA, Canada, Australia, Germany, UK and France), Defoort computed skilled emigration rates from 1975 to 2000 (one observation every five years). One can see that some regions experienced an increase in the intensity of the brain drain (especially Central America and sub-Saharan Africa) while significant decreases were observed in others (notably the Middle East and Northern Africa).

From Brain Drain to Brain Gain?

It is certainly a good thing for rich countries to host a skilled and talented workforce, and the move is also worthwhile (at least *ex ante)* from the perspective of the individual migrant. However, the social return to human capital is likely to exceed its private return given the many fiscal, technological, intra- and intergenerational (or Lucas-type) externalities involved. This externality argument is central in the early brain drain economic literature (Bhagwati and Hamada 1974), which emphasized that the brain drain entails significant losses for those left behind and contributes to increased inequality at the world level. Another negative aspect of the brain drain is that it can induce shortages of manpower in certain activities, for example when engineers or health professionals emigrate in disproportionately large numbers, thus undermining the ability of the origin country to adopt new technologies or deal with health crises. This can be reinforced by governments distorting the provision of public education away from general (portable) skills when graduates leave the country, with the country ending up educating too few nurses, doctors or engineers, and too many lawyers (Poutvaara 2004). The argument, however, can be reversed, since the prospect for migration may create a bias in the opposite direction (see Lucas 2005, for an illuminating analysis of the Philippines higher-education market).

The prospect of migration can also impact on the very decision as to whether to study. When education is a passport to emigration, migration prospects create additional incentives to invest in human capital. If migration is probabilistic in that people are uncertain about their chances of future migration when they make education decisions, then the incentive effect just described may more than compensate the brain drain effect, resulting in a higher level of human capital in the source country. As demonstrated in a series of recent

papers (for example, Mountford 1997; Beine et al. 2001), such a positive outcome is theoretically more likely when inter-country wage differentials are large enough to generate a high incentive effect and skilled emigration rates are sufficiently low. These theories have been confirmed empirically by Beine et al. (2007b), who found a positive and significant effect of migration prospects on human capital formation in a cross-section of 127 developing countries. From the latters' perspective, however, what matters is not how many of their native-born engage in higher education, but how many remain at home. To estimate the net effects country by country, Beine et al. (2007b) used counterfactual macro-simulations and found that countries combining relatively low levels of human capital and low skilled emigration rates are likely to experience a net gain. Their results show a positive effect on aggregate, but with more losers (which tend to lose a lot in relative terms) than winners. The situation of many small African and Central American countries appears extremely worrisome while the main globalizers (for example, India, China) all register moderate gains.

Feedback Effects

Remittances

The literature on migrants' remittances shows that the two main motivations to remit are altruism, on the one hand, and exchange, on the other hand (Rapoport and Docquier 2006). Altruism is primarily directed towards the immediate family, while remittances motivated by exchange pay for services such as care of the migrant's assets or relatives at home. Exchange-motivated transfers are typically observed in case of a temporary migration and signal the migrants' intention to return. It is therefore a priori unclear whether educated migrants remit more than their uneducated compatriots; the former may remit more to meet their implicit commitment to reimburse the family for funding of education investments (and, in addition, they have a higher income potential), but on the other hand, they tend to emigrate with their families, and on a more permanent basis. Indeed, at an aggregate level, Faini (2006) finds that brain drain migration (as measured by the proportion of skilled among emigrants) is associated with lower remittance inflows.

Return Migration and Brain Circulation

Return migration is rare among the highly educated unless sustained growth precedes return. For example, less than one-fifth of Taiwanese and Korean Ph.D. students who graduated from US universities in the 1970s in the fields of science and engineering returned to Taiwan or Korea, a proportion that rose to two-thirds in the course of the 1990s, after two decades of impressive growth in these countries. The figures for Chinese and Indian Ph.D. students graduating from US universities in the same fields during the 1990s are similar to those for Taiwan or Korea in the 1980s (OECD 2002). These numbers suggest that return skilled migration is more a consequence than a trigger of growth. On a more reduced scale, however, there are many case- studies showing clear signs of brain circulation. For example, a recent survey conducted among 225 Indian software firms concluded that 30–40 per cent of the higher-level employees had previous work experience in similar occupations in a developed country (Commander et al. 2004).

Diaspora Externalities

A large sociological literature emphasizes the potential for skilled migrants to reduce transaction and other types of information costs and thus facilitate trade, foreign direct investment (FDI) flows and technology transfers between their host and home countries. This has first been confirmed in the field of international trade (Gould 1994; Head and Ries 1998; Rauch and Casella 2003). Regarding FDI, Kugler and Rapoport (2007) used US data on immigration and FDI outflows and found that past skilled immigration significantly increases a country's chances of attracting FDI in the subsequent period. These results complement recent case studies of the software industry showing that skilled migrants take an active part in the creation of business networks that lead to FDI deployment in their home country (Arora and Gambardella 2005).

Conclusion

The number of skilled migrants from poor to rich countries has increased dramatically since the 1970s. In the face of rising wage differentials and of diverging demographic structures between rich and poor countries, this tendency is likely to be confirmed in the future. While the brain drain has long been viewed as detrimental to poor countries' growth potential, recent economic research has emphasized that, alongside positive feedback effects arising from skilled migrants' participation in business networks, one also has to consider the effect of migration prospects on human capital-building in source countries. This new literature suggests that a limited degree of skilled emigration could be beneficial for growth and development. Empirical research shows that this is indeed the case for a limited number of large, intermediate-income developing countries. For the vast majority of poor and small developing countries, however, current skilled emigration rates are most certainly well beyond any sustainable threshold level of brain drain.

Bibliography

Arora, A., and A. Gambardella. 2005. *From underdogs to tigers: The rise and growth of the software industry in Brazil, China, India, Ireland, and Israel*. Oxford/New York: Oxford University Press.

Beine, M., F. Docquier, and H. Rapoport. 2001. Brain drain and economic growth: Theory and evidence. *Journal of Development Economics* 64: 275–289.

Beine, M., Docquier, F. and Rapoport, H. 2007a. Measuring international skilled migration: A new database controlling for age of entry. *World Bank Economic Review* (forthcoming).

Beine, M., Docquier, F. and Rapoport, H. 2007b. Brain drain and human capital formation in developing countries: Winners and losers. *Economic Journal* (forthcoming).

Bhagwati, J.N., and K. Hamada. 1974. The brain drain: International integration of markets for professionals and unemployment. *Journal of Development Economics* 1: 19–42.

Carrington, W.J. and Detragiache, E. 1998. How big is the brain drain? Working Paper No. 98–102. Washington, DC: International Monetary Fund.

Commander, S., Chanda, R., Kangasniemi, M. and Winters, L.A. 2004. Must skilled migration be a brain drain? Evidence from the Indian software industry. Discussion Paper No. 1422. Bonn: Institute for the Study of Labor (IZA).

Defoort, C. 2006. Tendances de long terme en migrations internationales: analyse à partir de 6 pays receveurs. Mimeo, EQUIPPE, Universités de Lille, and IRES, Université Catholique de Louvain.

Docquier, F., and A. Marfouk. 2006. International migration by educational attainment (1990–2000). In *International migration, remittances and the brain drain*, ed. C. Ozden and M. Schiff. Basingstoke: Palgrave Macmillan.

Faini, R. 2006. Remittances and the brain drain. Discussion Paper No. 2155. Bonn: Institute for the Study of Labor (IZA).

Gould, D.M. 1994. Immigrant links to the home country: Empirical implications for U.S. bilateral trade flows. *Review of Economics and Statistics* 76: 302–316.

Head, K., and J. Ries. 1998. Immigration and trade creation: Econometric evidence from Canada. *Canadian Journal of Economics* 31: 47–62.

Kugler, M., and H. Rapoport. 2007. International labor and capital flows: Complements or substitutes? *Economics Letters* 94: 155–162.

Lucas, R.E.B. 2005. *International migration and economic development: Lessons from low-income countries*. Cheltenham: Edward Elgar.

Mountford, A. 1997. Can a brain drain be good for growth in the source economy? *Journal of Development Economics* 53: 287–303.

OECD (Organisation for Economic Co-operation and Development). 2002. *Trends in international migration*. Paris: OECD Editions.

Poutvaara, P. 2004. Public education in an integrated Europe: Studying to migrate and teaching to stay. Working Paper No. 1369. Munich: CESifo.

Rapoport, H., and F. Docquier. 2006. The economics of migrants' remittances. In *Handbook of the economics of giving, altruism and reciprocity*, vol. 2, ed. S.-C. Kolm and J. Mercier Ythier. Amsterdam: North-Holland.

Rauch, J.E., and A. Casella. 2003. Overcoming informational barriers to international resource allocation: Prices and ties. *Economic Journal* 113: 21–42.

Rosenzweig, M.R. 2005. Consequences of migration for developing countries. Paper prepared for the UN conference on international migration and development, Population Division.

Braudel, Fernand (1902–1985)

R. Forster

Keywords

Annales School; Braudel, F.; Demography; Geography; Technology

JEL Classifications
B31

One of the foremost social and economic historians of the 20th century, Fernand Braudel combined a perceptive grasp of historical interconnections, an exceptional skill of synthesis and an evocative, even 'poetic' style. Perception, scope and style were brought to successful fruition in Braudel's *La Méditerranée et le monde méditerranéen à l'époque de Phillipe II* (1949), which became a classic in historical literature and a model for a major school of French history known as the *Annales.* In this seminal volume and in many methodological articles that followed, Braudel proposed a triple notion of historical time – the long run (*longue durée*) over a millennium, trends (*conjunctures*) of a generation or more, and events (*événements*). According to Braudel, each of these notions or blocks of time involved unique historical problems, appropriate source materials, and even special approaches employing social-science disciplines neighbouring to history. Braudel's model emphasized the 'constraints' of human endeavour rather than the 'permissive' factors that had been so much a part of Whig history as practised by most early 20th-century historians. These constraints were imposed by geography, climate and soils, by demographic pressure, and by a static social structure held together by the bonds of custom. Braudel likened this 'structure' to a glacier or to the sea depths, imparting both a physical metaphor and a sense of timelessness or immobility. His second temporal level, the *conjoncture,* made some room for change as new technologies, new forms of economic organization (especially capitalism), and subtle shifts in social relations and customs altered the 'structure'. Braudel likened these changes – he preferred the term 'mutations' – to the sea tides. Finally the 'event' was a kind of surface noise, an indication perhaps of deeper sea changes, but in itself of little significance for the historian. He likened these events to whitecaps on the vast ocean.

In addition to his emphasis on constraints and the obligation of historians to understand their deterministic effects on human behaviour, Braudel also stressed the cyclical nature of most of history – 'le temps, quasi immobile, fait de repétitions, de retours insistants, de cycles sans cesse recommencés'. There was about Braudel a strong sense of romantic conservatism that challenged Marxist and Whig historian alike. Braudel imparted to the Annales School a preference for metaphors taken from biology and anthropology (interconnection, *liens*, mutations, *glissements*) instead of the vocabulary, and indeed the goals, of physics or economics (parsimonious cause, leanness of argument, elegance of formula or theory). It is also clear that for Braudel geography and demography were basic objects of study, that technology and economic and social organization were important, but that political history, biography and the history of formal ideas were secondary and even trivial historical pursuits. In a direct attack on the kind of history taught at the Sorbonne, Braudel insisted that 'events' tell us little about the deeper and interlocking structures and their subtle mutations. Indeed, such surface history may suggest a misguided 'voluntarism' in human history. With such a perspective, it is understandable that Braudel was most comfortable in the thousands of years of pre-industrial history. The more recent 19th century and its urban-industrial dynamism were unsettling to his outlook, his methodology and even to his aesthetic sense. But, like a cultural anthropologist, Braudel never ceased to stress the fact that most of world history was pre-industrial.

Although Braudel was interested in quantification, he was never a model-builder, and in fact he used numbers illustratively rather than systematically. He had much to do with the Annales-style deployment of an array of graphic techniques – often very artfully designed – to demonstrate proportions and relationships, but as a descriptive technique in which the reader had to access the results by eye. Braudel did not use statistical measures, much less economic theory, perhaps because he considered them too abstract, and a threat to the living texture of social history that was his main concern. In the 1970s, like much of the Annales School, Braudel moved further towards cultural anthropology as reflected in his

notion of 'day-to-dayness' (*la vie quotidienne*), in the cultural determinants of economic and social behaviour, in the values and attitudes (mentalities) of social groups, and in the *gestes* and *code* of an entire society or even a 'civilization'. These features were already present in the *Méditerranée*, but they became even more pronounced in his more recent *Civilisation matérielle et capitalisme* (*XV-XVII*^{e} *siècle*) (1967–1979).

Fernand Braudel was also director of the Maison des Sciences de l'Homme in Paris, professor at the Ecole Pratique des Hautes Etudes and at the Collège de France, and co-editor of the *Annales*: *ESC*, one of the most prestigious journals of social and economic history in the Western world today. Braudel's seminal writings, his provocative teaching, his administrative and editorial talents, and, not least, his powerful personality made him an 'animateur' of the 'School of the Annales' for more than 30 years. Yet his work stands on its own as an appeal to approach history in its widest scope in time and place (*histoire totale*), in alliance with neighbouring disciplines, and presented with that special verve we call 'Braudelian'.

Selected Works

1949, 1966. La Méditerranée et le monde méditerranéen à l'époque de Phillipe II. *Paris: Armand Colin. Trans. by S. Reynolds as* The Mediterranean and the Mediterranean World in the Age of Philip II*, London; Collins; New York: Harper & Row, 1972–3.*

1967, 1979. *Civilization matérielle et capitalisme, XV–XVII*^{e} *siècle.* 1967 edn, 1 vol.; 1979 edn, 3 vols; Paris: Armand Colin. Trans. of 1st edn by M. Kochan as *Capitalism and Material Life*, *1400–1800,* London: Weidenfeld & Nicolson, 1973. Trans. of 2nd edn by M. Kochan and S. Reynolds as *Civilization and Capitalism, 15th–18th Century,* London: Collins; New York: Harper & Row, 3 vols, 1981, 1982, 1984.

1969. *Ecrits sur l'histoire*. Paris: Flammarion.

1973. *Mélanges en l'honneur de F. Braudel.* 2 vols. Toulouse: Edouard Privat.

Braverman, Harry (1920–1976)

David M. Gordon

B

Keywords

Braverman, H.; Capitalism; Deskilling; Division of labour; Exploitation; Heilbroner, R.; Labour power; Marx's analysis of capitalist production; Reskilling; Supervision

JEL Classifications

B31

Harry Braverman was born in 1920 in New York City and died on 2 August 1976 in Honesdale, Pennsylvania.

Born into a working-class family, he was able to spend only one year in college before financial problems forced him out of Brooklyn College and into the Brooklyn Navy Yard. He worked there for eight years primarily as a coppersmith and then moved around the United States, working in the steel industry and in a variety of skilled trades. He became deeply involved in the trade union and socialist political movements. He helped found *The American Socialist* in 1954 and worked as its coeditor for five years. After the journal ceased publication for practical reasons, he moved into publishing, working first at Grove Press as an editor and eventually as vice-president and general business manager. In 1967 he became Managing Director of Monthly Review Press, where he worked until his death.

Braverman is best known for his classic study of the labour process under capitalism, *Labor and Monopoly Capital* (1974), awarded the 1974 C. Wright Mills Award. 'Until the appearance of Harry Braverman's remarkable book', Robert L. Heilbroner wrote in the *New York Review of Books*, 'there has been no broad view of the labour process as a whole...' The book was all the more remarkable because of the void it filled in the Marxian analytic tradition – a literature ostensibly grounded in the analysis of the structural effects of

class conflict but persistently reticent about the actual structure and experience of work in capitalist production.

Labour and Monopoly Capital advances three principal hypotheses about the labour process in capitalist societies.

First, Braverman helps formalize and extend Marx's resonant analysis, in Volume I of *Capital*, of the distinction between labour and labour power. Braverman highlights the essential importance and persistence of managerial efforts to gain increasing control over the labour process in order to rationalize – to render more predictable – the extraction of labour activity from productive employees.

Second, Braverman argues that such managerial efforts lead inevitably to the homogenization of work tasks and the reduction of skill required in productive jobs. He concludes (p. 83) that 'this might even be called the general law of the capitalist division of labor. It is not the sole force acting upon the organization of work, but it is certainly the most powerful and general.'

Third, as a corollary of the second hypothesis, Braverman argues both analytically and with rich empirical detail that this 'general law of the capitalist division of labour' applies just as clearly to later stages of capitalist development, with their proliferation of office jobs and white collars, as to the earlier stages of competitive capitalism and largely industrial work.

The first analytic strand of Braverman's work was both seminal and crucial in helping foster a renaissance of Marxian analyses of the labour process. The second and third hypotheses have proved more controversial. There are two grounds for concern. Braverman's analysis tends to reduce the character of the labour process to essentially one dimension – the level of skill required and control permitted by embodied skills – and therefore unnecessarily compresses the *many* essential dimensions of worker activity and effectiveness in production to a single monotonic index. At the same time, there is good reason for worrying about the simplicity of Braverman's argument of historically irreversible 'deskilling' for all segments of the productive working class; it is quite plausible to hypothesize that for some labour segments in recent phases of capitalist development there has been a 'reskilling', as many have since called it, which has not in any way liberated these workers from capitalist exploitation or intensive managerial supervision.

Selected Works

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

1976. Two comments. Special issue on 'Technology, the Labor Process, and the Working Class'. *Monthly Review* 28(3): 119–124.

Bray, John (1809–1897)

N. W. Thompson

John Bray was born in the United States but spent his formative years (1822–1842) in England. His attention was drawn to social and industrial questions during a period as an itinerant printer in the early 1830s and also through his work with the unstamped *Voice of the West Riding* (1833–1834).

In 1837 Bray gave a series of lectures to the Leeds Working Men's Association – lectures which were to form the basis of his one major work *Labour's Wrongs and Labour's Remedy*, published in 1839. Shortly after (1842) he emigrated to the United States. However, his letters to the American papers show that he remained concerned with social and political matters as they touched upon the interests of the labouring-classes; indeed, in the 1880s he became involved with the syndicalist Knights of Labor and was hailed in 1885, by the *Detroit News*, as the oldest living socialist born in America.

In *Labour's Wrongs* Bray traced the impoverishment of the labouring-classes to the skewed distribution of the ownership of the nation's productive capacity, which permitted the coercive exercise of economic power by the few against

the interests of the many. This power was used to exploit those with only their labour to sell by means of unequal exchanges that reduced the value of labour to a bare subsistence level. Thus for Bray it was more by the infraction of the principle of equal exchanges 'by the capitalist, than by all other causes united that inequality of condition is produced and maintained and the working man offered up bound hand and foot, a sacrifice upon the altar of Mammon'. For Bray, therefore, exploitation occurred in the sphere of exchange with the crucial intermediation of money, which he saw as instrumental in ensuring that everything 'generated by the power of labour is perpetually carried off and absorbed by capital'. Further, the impoverishment of labour that resulted caused deficient demand and, in consequence, general economic depression.

Bray's solution to the iniquities and inequities of competitive capitalism was the creation of an economic system that would guarantee 'universal labour and equal exchanges'. Like other nineteenth-century socialist and anti-capitalist writers Bray sought to transmute the labour theory of value from a critical tool to an operational imperative. Thus goods should exchange at their labour values, for with labour exchanged against labour: 'That which is now called profit and interest cannot exist.' This was to be achieved by ensuring that the means of production were 'possessed and controlled by society at large' – something which was to be secured through purchase, the purchase price being met out of wealth created once the nation's productive capacity was under collective control.

Bray does no more than sketch the operational outlines of this socialist commonwealth, but it is clear that although influenced by Owenite thinking, his conception of socialism involved a move away from the idea of self-contained, self-sufficient, cooperative communities in the direction of central control over output, pricing, allocation and distribution. In this respect, while bearing many of the hallmarks of early nineteenth-century Owenite socialism, *Labour's Wrongs* points to the work of late nineteenth-century socialists where the market is supplanted by planning.

Selected Works

1839. *Labour's wrongs and labour's remedy or, the age of might and the age of right*. Leeds. Reprinted, New York: A.M. Kelley, 1968.

References

Beales, H.L. 1933. *The early English socialists*. London: Hamish Hamilton.

Beer, M. 1953. *A history of British socialism*, 2 vols. London: Allen & Unwin.

Carr, H.J. 1940. John Francis Bray. *Economica* 7: 397–415.

Cole, G.D.H. 1977. *A history of socialist thought*, Socialist Thought: the Forerunners, 1789–1850, vol. 1. London: Macmillan.

Foxwell, H.S. 1899. Introduction to the English translation of A. Menger. *The right to the whole produce of labour*. London: Macmillan.

Gray, A. 1967. *The socialist tradition: Moses to Lenin*. London: Longman.

Henderson, J.P. 1985. An English communist, Mr Bray (and) his remarkable work. *History of Political Economy* 17: 73–95.

Hunt, E.K. 1980. The relation of the Ricardian socialists to Ricardo and Marx. *Science and Society* 44: 177–198.

King, J.E. 1981. Perish Commerce! Free trade and underconsumption in early British radical economics. *Australian Economic Papers* 20: 235–257.

King, J.E. 1983. Utopian or scientific? A reconsideration of the Ricardian socialists. *History of Political Economy* 15: 345–373.

Lloyd-Prichard, M.F. 1957. Introduction to J.F. Bray. *A voyage from Utopia*. London: Lawrence & Wishart.

Lowenthal, E. 1911. *The Ricardian socialists*. New York: Longman.

Thompson, N.W. 1984. *The people's science: The popular political economy of exploitation and crisis, 1816–34*. Cambridge: Cambridge University Press.

Breckinridge, Sophonisba Preston (1866–1948)

B. Berch

Abstract

Born on 1 April 1866 in Lexington, Kentucky; died on 30 July 1948 in Chicago, Illinois. Breckinridge (Wellesley '88), the first woman

to pass the bar examination in Kentucky, abandoned legal work to take a PhD in political science at the University of Chicago, which she completed in 1901, followed by a law degree in 1904. Part of the circle of social reformers centred around Jane Addams at Hull House, Breckinridge pioneered in the professionalization of social work (as teacher, then as Dean and head of research of the Chicago School of Civics and Philanthropy, where social workers were trained). Her methodology was radically empirical; social problems were to be studied in their concrete context, by first-hand observation of the homes and communities of the poor. Working closely with Edith Abbott, she produced numerous monographs on tenement life and the effects of urban poverty on the breakdown of families. *New Homes for Old* (1921) detailed the dislocations and privations of the immigrant poor in big cities, while giving the social worker the leading role in helping these hapless victims construct a decent life. As early as the 1920s, Breckinridge was emphasizing the need for government responsibility for social welfare programmes, an idea not popular in America until the Depression of the 1930s. In 1927 she helped found the *Social Service Review* which she edited for the rest of her life.

Born on 1 April 1866 in Lexington, Kentucky; died on 30 July 1948 in Chicago, Illinois. Breckinridge (Wellesley '88), the first woman to pass the bar examination in Kentucky, abandoned legal work to take a PhD in political science at the University of Chicago, which she completed in 1901, followed by a law degree in 1904. Part of the circle of social reformers centred around Jane Addams at Hull House, Breckinridge pioneered in the professionalization of social work (as teacher, then as Dean and head of research of the Chicago School of Civics and Philanthropy, where social workers were trained). Her methodology was radically empirical; social problems were to be studied in their concrete context, by first-hand observation of the homes and communities of the poor. Working closely with Edith Abbott, she produced numerous monographs on tenement life and the effects of urban poverty on the breakdown of families. *New Homes for Old* (1921) detailed the dislocations and privations of the immigrant poor in big cities, while giving the social worker the leading role in helping these hapless victims construct a decent life. As early as the 1920s, Breckinridge was emphasizing the need for government responsibility for social welfare programmes, an idea not popular in America until the Depression of the 1930s. In 1927 she helped found the *Social Service Review* which she edited for the rest of her life.

Selected Works

1912. (With E. Abbott.) *The delinquent child and the home*. New York: Charities Publications Committee, Russell Sage Foundation.
1921. *New homes for old*. New York: Harper & Brothers.
1924. *Family welfare work in a metropolitan community*. Chicago: University of Chicago Press.
1927. *Public welfare administration in the United States*. Chicago: University of Chicago Press.
1931. *Marriage and the civic rights of women*. Chicago: University of Chicago Press.
1933. *Women in the twentieth century*. New York: McGraw-Hill.
1934. *The family and the state*. Chicago: University of Chicago Press.

Brentano, Lujo (Ludwig Josef) (1844–1931)

B. Schefold

Keywords

Brentano, L.; Economic history; Engel, E.; Great Depression; Malthus's theory of population; Marx, K. H.; Profit sharing; Subjective theory of value; Tariffs; Trade unions

JEL Classifications
B31

Brentano was born in Aschaffenburg (Germany) into an old patrician family. Clemens Brentano, the poet, was his uncle; Bettina von Arnim, the writer, his aunt; and Franz Brentano, the philosopher, his brother. He was brought up in an atmosphere dominated by Catholicism (which he was later to abandon after the declaration of papal infallibility) and was particularly influenced by the anti-Prussian tradition of southern Germany. He studied law and economics in Heidelberg and Göttingen. From 1871 he taught political economy as professor in Berlin, Breslau, Strassburg, Vienna, Leipzig and Munich.

A decisive point for his later career was his participation in the Statistical Seminar connected with the Prussian Statistical Office. Its director was Ernst Engel (originator of Engel's law), whose strong interest in the social conditions of the working classes was to have a lasting influence on Brentano. Engel advocated profit-sharing schemes as a means to the solution of the social question. In 1868 Brentano accompanied him on a visit to England, where they studied the effects of such measures. His experiences in England convinced Brentano of the inadequacy of profit-sharing for the reform of capitalism, but suggested another approach, which was to remain the main topic of Brentano's intellectual work: the improvement of the worker's position in the labour market through the establishment of trade unions.

While the individual worker was forced to sell his labour power under any conditions, this would not be the case for an organized coalition of workers. Such a coalition would enable them to become as free and independent as the sellers of other commodities and would allow for an effective control of the labour supply (1871–2, vol. 2; 1877, ch. 2). It was Brentano's deep conviction that trade unions were the only means to secure an adequate participation of the working classes in the general increase of wealth. He was especially interested in the history of the trade unions, which he traced back to the medieval guilds (1871–2, vol. 1). Especially interesting – particularly for the current debate – was his discussion of positive productivity effects of labour time reductions (1876).

He regarded the introduction of a general social security system as another important step for the reform of capitalism. He also favoured the cartelization of Germany industry. It was characteristic of him that he always intended to solve the social question within the framework of a capitalist economic system. He therefore rejected Marx and the Social Democrats of 19th-century Germany. Brentano emphasized that unequal conditions of material existence were absolutely necessary for the further cultural advancement of mankind (1877, pp. 303–4).

His concern for the social question shaped Brentano's attitude towards the classical economists: he opposed the classical notion of an abstract profit-maximizing individual as the central axiom of political economy, and found this particularly inadequate to describe working-class behaviour and the labour market (1923, ch. 1). It is in this context that his preoccupation with economic history (1916; 1927–9) must be seen. He intended to show that the relations between man and the economic system were changing through history, and that the individual of classical economics was not the starting-point, but the result of economic development (1927–9, vol. 1, pp. iii–iv).

Further fields of interest were Malthus's theory of population development (1924), the theory of value (where he favoured the subjective theory of value; 1924), the German corn tariffs (which he opposed), and different forms of the law of estate.

Throughout his life Brentano remained an open-minded and enlightened liberal of whom an English trade union leader once said: 'He was our friend before it was fashionable to be our friend.' Brentano was a founding member of the Verein für Socialpolitik, which he left in 1929, when he thought that it had become reactionary. He opposed Bismarck in the Kaiserreich, the extreme German annexationists during the First World War – although himself favouring limited territorial expansion – and the Socialist Revolutionaries in the post-war period. The republican

government considered his appointment as first German post-war ambassador to Washington, but because of his advanced age he declined.

During the Weimar Republic Brentano was still concerned with social policy, mainly with the struggle for the eight-hour working day. He deplored the harsh austerity policy during the Great Depression. His memoirs, written in 1930, ended: 'I do not understand this policy. Do they want a social revolution?' (1931, p. 404).

Selected Works

1870. *On the history and development of guilds, and the origin of trade unions.* London: Trübner.

1871–2. *Die Arbeitergilden in der Gegenwart. Vol. 1: Zur Geschichte der englischen Gewerkvereine.* Leipzig: Duncker & Humblot, 1871; *Vol. 2: Zur Kritik der englischen Gewerkvereine.* Leipzig: Duncker & Humblot, 1872.

1876. *Über das Verhältnis von Arbeitslohn und Arbeitszeit zur Arbeitsleistung.* Leipzig: Duncker & Humblot. Trans. as *Hours and wages in relation to production.* New York: Scribner; London: Sonnenschein, 1894.

1877. *Das Arbeitsverhältnis gemäss dem heutigen Recht.* Leipzig: Duncker & Humblot. *Trans. as The relation of labour to the law of to-day.* New York: Putnam, 1898.

1889. *Über die Ursachen der heutigen socialen Noth.* Leipzig: Duncker & Humblot.

1916. *Die Anfänge des modernen Kapitalismus.* Munich: Königlich Bayerische Akademie der Wissenschaften.

1923. *Der wirtschaftende Mensch in der Geschichte.* Leipzig: Felix Meiner.

1924. *Konkrete Grundbedingungen der Volkswirtschaft.* Leipzig: Felix Meiner.

1927–9. *Eine Geschichte der wirtschaftlichen Entwicklung Englands. 3 Vols.* Jena: Gustav Fischer.

1931. *Mein Leben im Kampf um die soziale Entwicklung Deutschlands*. Jena: Eugen Diederichs.

Bresciani-Turroni, Costantino (1882–1963)

Henry W. Spiegel

Keywords

Bortkiewicz, L. von; Bresciasni-Turroni, C.; *Deflation*; Forecasting; Foreign exchange equilibrium; Galiani, F.; German hyperinflation; Great Depression; Loria, A.; Pantaleoni, M.; Pareto's law of income distribution; Productivity theory of interest; Quantity theory of money; Ricca-Salerno, G.; Robinson, J. V.; Schmoller, G. von; Speculation; Stabilization policy; Wage indexation; Wagner, A.

JEL Classifications

B31

The last great exponent of old-time liberalism in Italian economics, Bresciani was an Italian counterpart of such distinguished libertarians as Robbins, Hayek or Friedman, a bit more moderate, perhaps, in his views and with a quantitative bent at least equal to Friedman's. Bresciani was born in Verona and his teachers in his homeland included Ricca-Salerno and Loria. After the completion of his studies at a number of universities in Italy, he went to the University of Berlin, at that time at the height of its prestige, to study with historical economists such as Adolf Wagner and Gustav Schmoller, and with L. von Bortkiewicz, the mathematical statistician and pioneer in Marxian econometrics.

Amidst the push and pull of these intellectual influences, Bresciani preserved an admirable independence of mind. Loria did not convert him to socialism and Schmoller did not turn him into an historical economist. More influenced by Pareto and Pantaleoni than by his great teachers, he became, first of all, an economic theorist, but again not a pure one but one looking for statistical verifications of theoretical propositions.

In his writings he would give a respectful hearing to the views of the classics and provide copious

references to modern authorities, foreign languages and mathematical modes of expression constituting no barriers. As an Italian and libertarian, he was especially fond of citing Galiani. After the publication of Keynes's *General Theory* in 1936, Bresciani, like other contemporary economists, had to come to terms with the new economics. Again he showed his independence by continuing to adhere to such established doctrines as the quantity theory of money and the productivity theory of interest. This attitude, together with his insistence on the limitations rather than opportunities of public policies, gave an old-fashioned flavour to his later writings, published, as they were, at a time when Keynes's influence reached its peak.

Bresciani's teaching career, which included chairs in statistics, led him eventually to the University of Milan (1926–57), but his work there was interrupted by various other activities. During the 1920s he served as an adviser to the Berlin office of the Allied Reparations Commission, and from 1927 to 1940 he lectured at the newly established Egyptian University of Cairo. This multiplication of jobs again confirmed his penchant for independence and gave him the opportunity to absent himself from fascist Italy. After the Second World War he served the new Republic of Italy as president of an important bank and for a brief period also as minister of foreign trade. In this capacity he again demonstrated his independence, this time from ideological preferences, by sponsoring a government organization for export credit and insurance.

As a writer Bresciani started out, at age 22, with a critical review of Pareto's law of income distribution, a subject to which he returned later more than once. Much of his work was devoted to the theory of prices, domestic and international, present and future, as well as the relation between prices and interest. Among other topics that he investigated were the influence of speculation on prices, which he recognized as not always beneficial, economic forecasting, the inductive verification of the theory of international payments, and the relation between the harvest and the price of cotton in Egypt. Late in life he wrote a number of broad syntheses of economics, including a two-volume *Corso* that went into many editions.

Bresciani's masterpiece, and the work for which he is best known, is *The Economics of Inflation,* published originally in Italian in 1931 and in a revised English translation in 1937. The Italian title of the book – *Le vicende del marco tedesco,* or the vicissitudes of the German mark – conveys the substance of the book better than the title of the English translation, which claims a level of abstraction far higher than that embodied in the work, and, correspondingly, a much wider applicability of the content. The subtitle of the English translation is also carelessly worded. The subject of the work is the great German inflation after the First World War, when prices had risen to astronomical heights and $1 in the end purchased 42 marks followed by 11 zeros. At that time this was considered a record, but the Hungarian inflation after the Second World War surpassed it, with the dollar then buying 145 pengö followed by 27 zeros.

Bresciani's book has been the standard work on the subject ever since. What was open to debate was never the completeness or reliability of the material that he presented but his interpretation. German students of the matter tended to adhere to the view that the rise in prices reflected the unfavourable rate of exchange, which in turn was ascribed, at least in part, to the burden of reparation payments that the Germans were eager to demonstrate as outrageously unreasonable. Bresciani opposed this interpretation. His principal argument was that foreign exchanges, by means of well-known mechanisms, will never fail to reach an equilibrium if only the external value of the currency falls deeply enough. Bresciani, instead of putting the blame on the foreign exchanges, placed it firmly on the German authorities which pursued policies of fiscal irresponsibility and unrestrained monetary expansion. Bresciani also discussed still other interpretations – conspiratorial or scandal theories – but found them unconvincing. One variant of these made the industrialists, who gained so much from the galloping inflation, responsible for it. Another one put the onus on the German authorities' desire to prove the

impossibility of reparation payments. It may be of some interest that the second variant of the scandal theory would constitute a corollary of the policy of *deflation* which Chancellor Brüning adopted a few years later during the Great Depression, a policy instrumental in helping Germany to rid herself of reparation payments.

Critics of the work brought still other points of view before the reader. Joan Robinson, to give an example, stressed the role of ever-rising money wages that became indexed and subject to automatic increases. This would seem to lend support to the view blaming the foreign exchanges, because the rise in money wages offset the forces making for equilibrium of the foreign-exchange rates. But Robinson does not fully endorse Bresciani's or the German interpretation. In her view the eventual stabilization of the mark in November 1923 does not support the conclusion that monetary stringency is necessary and sufficient to put an end to inflation. In Robinson's view the stabilization succeeded because by that time the old German mark had shed almost all the standard functions that money is to serve.

Selected Works

1929. The movement of wages in Germany during the depreciation of the mark and after stabilization. *Journal of the Royal Statistical Society* 92: 374–414.

1931. *Le vicenda del marco tedesco*. Trans. by M. Sayers as *The economics of inflation: A study of currency depreciation in post-war Germany 1914–1923,* with a foreword by Lionel Robbins. London/New York: Allen & Unwin/Barnes & Noble, 1937.

1932. *Inductive verification of the theory of international payments*. *Cairo: Noury.*

1934a. Egypt's balance of trade. *Journal of Political Economy* 42: 371–384.

1934b. The 'purchasing power parity' doctrine. *L'Egypte Contemporaine* 25: 433–464.

1936. The theory of saving. *Economica* 3, Part I, February, 1–23, Part II, May, 162–181.

1937. *On Pareto's Law. Journal of the Royal Statistical Society 100: 421–432.*

1938. The multiplier in practice: Some results of recent German experience. *Review of Economics and Statistics* 20(May): 76–88.

1939. Annual survey of statistical data: Pareto's Law and the index of inequality of incomes. *Econometrica* 7(April): 107–133.

1942. Economic policy for the thinking man. *London: Hodge.*

1964. Articles contributed by Costantino Bresciani-Turroni to the *Review of economic conditions in Italy* in the years from 1947 to 1962. Rome: Banco di Roma.

Bibliography

Anon. 1964. Obituary of C. Bresciani-Turroni, with bibliography. In *Review of economic conditions in Italy*, vol. 3, Rome: Banco di Roma.

Gambino, A. 1965. La politica monetaria *e* creditizia negli scritti di C. Bresciani-Turroni. *Studi economici* 20: 196–219.

Gambino, A. 1972. *C. Bresciani-Turroni. Dizionario biografico degli italiani*. Vol. 14. Rome: Instituto della Enciclopedia Italiana.

Robinson, J. 1938. Review of the economics of inflation. *Economic Journal* 48: 507–513.

Tamagna, F.M. 1967. C. Bresciani-Turroni. In *International encyclopedia of the social sciences*, vol. 2. New York: Macmillan.

Bretton Woods System

Peter B. Kenen

Abstract

Under the Bretton Woods System, created after the Second World War, each country had to peg its currency to gold or to the US dollar, but it could obtain temporary financing from the International Monetary Fund. In practice, countries pegged their currencies to the dollar and accumulated dollar reserves, which they could use to buy gold from the US Treasury. This regime served to finance US payments deficits but prevented the United States from changing its exchange rate. The system was undermined when other countries' dollar

holdings came to exceed US gold holdings. It was abandoned in 1973, when the major industrial countries let their currencies float.

Keywords

Bancor; Bank of England; Bretton Woods System; Capital controls; Convertibility; Devaluation; Exchange rate stability; Fixed exchange rates; Floating exchange rates; Foreign-exchange market; Gold exchange standard; Gold standard; International Monetary Fund; Keynes, J. M; Nixon, R; Nurkse, R; Reserve currency; Special Drawing Rights; Trade unions; Triffin, R; White, H

JEL Classification

F3

The international monetary system established at the end of the Second World War is commonly known as the Bretton Woods System. It takes its name from the conference held at Bretton Woods, New Hampshire, USA, in 1944, which adopted the Articles of Agreement of the International Monetary Fund (IMF) and thus put in place the rules and arrangements that would govern international monetary relations in the post-war world.

A comprehensive history of the Bretton Woods System would have to review the monetary and fiscal policies of the major industrial countries, most notably those of the United States and United Kingdom, the key-currency countries, describe the evolution of monetary cooperation, and recite the history of the IMF itself. An analytic assessment would have to examine balance-of-payments adjustment under the Bretton Woods System and compare the merits of pegged and floating exchange rates.

This account has narrower objectives. It reviews the origins of the system, the rules adopted at Bretton Woods, the differences between those rules and the way the system worked in practice, and the forces leading to the breakdown of the system in the early 1970s. Readers who want more detailed accounts may consult Cooper (1968); Solomon (1982); de Vries (1987); James (1996); Eichengreen (2006), and the official histories of the IMF (Horsefield 1969; de Vries 1976, 1986).

The Origins of the System

The design of the Bretton Woods System cannot be understood without recalling the monetary history of the interwar period and the lessons drawn from it at the time. Recent writers have drawn somewhat different lessons. Thus, Eichengreen (1991) argues that the credibility of the gold standard in the decades before the First World War depended on close cooperation among central banks, not on the exercise of hegemonic influence by the Bank of England, and that the absence of comparable cooperation doomed the gold-standard arrangements of the interwar period; he also argues that fiscal rigidities greatly compounded the problems of monetary management. But these are lessons for our time, reflecting recent concerns, not those that influenced the design of the Bretton Woods System.

At the end of the First World War, governments were firmly committed to the restoration of the gold standard, and most of them returned to gold during the 1920s. They did so unilaterally and sequentially, however, by adopting gold values for their own currencies. Although some such as Keynes (1925) warned them of the risks they were running, they paid too little attention to the pattern of exchange rates established by their actions. Nor did they understand completely the new environment in which they would have to maintain the gold standard – how monetary and fiscal policies would be constrained by the transfer of financial activity and influence from London to New York, by the domestic and foreign debt-service burdens built up by wartime borrowing, and by the increased power of the trade unions and of the political parties affiliated with them.

The new gold standard collapsed in fewer than 10 years, in the same sequential way that it was put together. Country after country let go of gold and allowed its exchange rate to float – to be determined by supply and demand in the foreign-exchange market – but they soon began to intervene in that market in order to influence the

behaviour of exchange rates. Even at that point, moreover, they acted unilaterally, not cooperatively. Central banks began to cooperate in the late 1930s, but the process was halted by the outbreak of war and the imposition of currency controls.

What lessons were learned from this experience? Writing for the League of Nations (1944, p. 210), Ragnar Nurkse put them in terms that were widely endorsed at the time. The setting of exchange rates, he concluded, could not be left to market forces:

> A system of completely free and flexible exchange rates is conceivable and may have certain attractions in theory. . . . Yet nothing would be more at variance with the lessons of the past.
>
> Freely fluctuating exchanges involve three serious disadvantages. In the first place, they create an element of risk which tends to discourage international trade....
>
> Secondly, as a means of adjusting the balance of payments, exchange fluctuations involve constant shifts of labour and other resources between production for the home market and production for export. Such shifts may be
>
> costly . . . and are obviously wasteful if the exchange-market conditions that call for them are temporary....
>
> Thirdly, experience has shown that fluctuating exchanges cannot always be relied upon to promote adjustment. Any considerable or continuous movement of the exchange rate is liable to generate anticipations of a further movement in the same direction.

Yet the setting of exchange rates, Nurkse argued, cannot be left to individual governments:

> An exchange rate by definition concerns more currencies than one. Yet exchange stabilization [in the interwar period] was carried out as an act of national sovereignty in one country after another with little or no regard to the resulting interrelationship of currency values in comparison with cost and price levels. . . . The piecemeal and haphazard manner of international monetary reconstruction sowed the seeds of subsequent disintegration. (League of Nations 1944, pp. 116–17)

Finally, governments should not be expected to sacrifice domestic economic stability merely to maintain exchange rate stability:

> Experience has shown that stability of exchange rates can no longer be achieved by domestic income adjustments if these involve depression and unemployment. Nor can it be achieved if such income adjustments involve a general inflation of prices which the country concerned is not prepared to endure. It is therefore only as a consequence of internal stability . . . that there can be any hope of securing a satisfactory degree of exchange stability as well. (League of Nations 1944, p. 229)

The plans that governments drafted in anticipation of the Bretton Woods conference differed in many ways but did not disagree about these matters. A new international institution would be needed to supervise exchange rate policies, in order to promote exchange rate stability and prevent competitive devaluations, but it would also have to concern itself with 'the promotion and maintenance of high levels of employment and real income' (Articles of Agreement, Article I (ii)).

The Design of the System

The design of the new monetary system was decided before the Bretton Woods conference, in talks between British and American negotiators. The British were led by John Maynard Keynes, the Americans by Harry Dexter White, and the two countries' proposals are known as the Keynes and White plans. They differed mainly in the way that they would provide financing for external imbalances. (On the plans and subsequent negotiations, see Gardner 1969; Horsefield 1969; Dam 1982.)

The Keynes plan was quite radical and reflected Keynes's concerns about the post-war situation. In the short run, Britain would need balance-of-payments financing; in the long run, the United States was likely to experience another depression, driving other countries into balance-of-payments deficit, and forcing them to choose between domestic stability and exchange rate stability if they could not obtain adequate financing. Hence, Keynes sought to create a monetary institution able to issue a new international currency (which Keynes called 'bancor'); it would be held and used by governments and central banks for settling external imbalances.

The White plan was more conservative and reflected White's concern that a large and elastic supply of international money would give other

countries an open- ended claim on the real resources of the United States. (In other words, the United States would wind up holding all of Keynes's bancor.) Hence, White sought to limit the supply of reserve credit by providing the new monetary institution with a finite pool of national currencies and gold, rather than the power to issue a new money of its own. (Ironically, the White plan failed to anticipate the emergence of the US dollar as a reserve currency, which made the supply of reserves very elastic and helped to undermine the Bretton Woods System at the start of the 1970s, when it became apparent that the United States could not maintain convertibility between the dollar and gold.)

The plan adopted at Bretton Woods was much like the White plan, although it made concessions to Keynes's concern about the danger of a deep US depression. If a country's currency became 'scarce' in world trade and in the IMF itself, because the country was running a balance-of-payments surplus, the IMF could ration that currency and authorize its members to limit imports from the surplus country. (This clause was never invoked, however, even in the years of the so-called dollar shortage.)

The Bretton Woods System imposed two major obligations on national governments but gave them something in exchange.

First, every member of the IMF had to peg its currency to gold or the US dollar (which was, in turn, pegged to gold at $35 per ounce). The IMF had to approve the initial exchange rate and every significant change thereafter. Before it could change its exchange rate, moreover, a government would have to show that it faced a 'fundamental disequilibrium' in its external accounts. That term was not defined, however, and led to much debate. It came to be interpreted eventually as an unsustainable conflict between 'external' and 'internal' balance – a situation in which a country could not defend its exchange rate without suffering substantial unemployment or inflation; see Nurkse (1945) and Meade (1951). (The operational issues resemble those which still bedevil attempts to define a fundamental equilibrium exchange rate; see, for example, Williamson 1983a, and International Monetary Fund 1984.)

With one notable exception, namely Canada, the major industrial countries did peg their exchange rates until the end of the 1960s and did not change them often. There was an extensive exchange rate realignment in 1949, triggered by a devaluation of sterling, but only a handful of changes thereafter. When they did change their rates, however, they did not let the IMF exercise effective supervision; it was informed at the very last minute, too late to offer advice or object. Developing countries, by contrast, adopted many exchange rate arrangements; a few had freely floating rates, and some had separate rates for different classes of transactions, with some rates pegged and others floating.

Second, every member of the IMF was expected to make its currency convertible as soon as possible. It could continue to control capital movements; recall the view expressed by Nurkse, that capital flows had been destabilizing in the interwar years. It could likewise continue to use tariffs and other trade controls for commercial-policy purposes. But it could not keep the resident of another country from using or converting domestic currency acquired from a current-account transaction. A Dane who earned French francs from exports to France was free to use them for another current-account transaction, sell them to someone else wanting to use them, or sell them to the Danish National Bank, which could then present them to the Bank of France for conversion into Danish currency.

Britain made the pound *fully* convertible for foreigners in 1947, for capital as well as current-account purposes, but it had to retreat speedily when countries that had built up large sterling balances during the war rushed to exchange them for dollars and drained away a large part of a large US loan to Britain. Thereafter, most governments moved cautiously toward current-account convertibility. Western Europe did not reach it until 1958, and some European countries did not abolish all of their capital controls until 1990; see Triffin (1957) and Kaplan and Schleiminger (1989).

In exchange for these commitments, members of the IMF were entitled to draw on the Fund's holdings of currencies and gold when they ran

balance-of-payments deficits and could not finance them by drawing down their own reserves. Each IMF member was given a *quota* that governed its subscription to the currency pool, how much it could draw from the pool, and its voting power in the IMF.

The Articles of Agreement, however, did not spell out the conditions under which countries could draw on the pool, and this became a contentious issue. The United States maintained that strict policy conditions would safeguard prospects for repayment and thus protect the drawing rights of other members. Other governments maintained that access should be automatic when a member needed short-term financing. The United States won this battle too, however, and access to most of the Fund's resources was (and remains) tightly linked to policy commitments made in advance by the government involved and monitored closely by the Fund. (On the origins and evolution of IMF conditionality, see Horsefield 1969; for criticism from various perspectives, see Dell 1981; Williamson 1983b; Kenen 1986.)

The Functioning of the System

Under the Bretton Woods System, all governments had the same rights and obligations. But the monetary system did not function symmetrically. (For more on the asymmetries discussed below, see Cooper 1972; Whitman 1974.)

First, there was a basic asymmetry between the situations of surplus and deficit countries – an asymmetry typical of pegged-rate regimes. A country can run a balance-of-payments surplus forever, although it may become uncomfortable with the domestic monetary consequences. There is no upper limit to the stock of reserves that a surplus country can acquire when it intervenes in foreign-exchange markets to keep its currency from appreciating. But a country cannot run a deficit for ever. It will exhaust its reserves as it goes on intervening to keep its currency from depreciating. The speed at which it loses them, moreover, is likely to accelerate as its holdings fall; speculative pressures will build up as foreign-exchange markets become convinced that the country will have to devalue its currency. Therefore, pegged-rate regimes tend to display a devaluation bias.

This bias would not matter in a two-country world, where the devaluation of one currency is no different from a revaluation of the other. It matters importantly in a multi-country world, where devaluation by a deficit country revalues every other currency, not just the surplus countries' currencies, and revaluation by a surplus country devalues every other currency, not just the deficit countries' currencies. And the bias had significant effects on the viability of the Bretton Woods System.

Devaluations by deficit countries were more frequent than revaluations by surplus countries, causing a gradual revaluation of the US dollar that weakened the competitive position of the United States.

This effect could have been offset by a devaluation of the dollar, but other asymmetries made that difficult. The dominance of the US economy and the key- currency role of the US dollar conferred important privileges on the United States but also limited its policy options.

The size and comparative stability of the US economy made for an asymmetry in policy determination. For most of the life of the Bretton Woods System, US monetary and fiscal policies were aimed exclusively at domestic targets – high employment, economic growth and price stability. There was no true policy coordination between the United States and the other industrial countries, although there were frequent consultations, especially in the 1960s. There were instead one-sided adaptations by the other countries, as they sought to keep their economies in line with the US economy; see, for example, Artis and Ostry (1986) and Kenen (1989).

Furthermore, the strength of the US economy permitted the United States to forgo an active exchange rate policy until the final years of the Bretton Woods System. It was the 'n^{th} country' in the system, whose exchange rate reflected the exchange rate policies of all other countries.

The passivity of the United States was helpful from one standpoint. In a world with n countries and currencies, there are only $n - 1$ independent exchange rates, which makes it impossible for all n countries to pursue independent exchange rate

policies (Mundell 1969). Therefore, the passivity of the United States helped to avoid policy conflict. Nevertheless, the arrangements supporting and promoting that passivity made the Bretton Woods System too brittle, forcing the United States to take very damaging measures in 1971, when it tried to achieve an exchange-rate realignment. Most countries defined their exchange rates with reference to the dollar, not gold, and stabilized those rates by buying and selling dollars. Hence, it was unnecessary for the United States to stabilize the dollar by buying and selling other countries' currencies. But it was also impossible for the United States to conduct an exchange rate policy of its own without other countries' tacit consent. It could change the gold price of the dollar, but it could not change the Deutschemark, franc and yen prices if Germany, France and Japan refused to change the dollar prices of their national currencies.

These asymmetries led to others. The US dollar was the only important convertible currency at the end of the Second World War, which caused it to become the key currency of the Bretton Woods System. It was used for official intervention in the foreign-exchange market and held along with gold as a reserve asset. There was, indeed, a neat division of labour under the Bretton Woods System. By buying and selling dollars in foreign-exchange markets, other governments stabilized the value of the dollar in terms of their national currencies. For its part, the United States stood ready to swap gold for dollars at $35 per ounce, making gold and dollars nearly perfect substitutes for the holders of reserves.

This arrangement imparted elasticity to the supply of reserves. Other governments wanting additional reserves could accumulate dollars, rather than compete for limited supplies of gold. But it had two serious defects.

First, it allowed the United States to run balance-of-payments deficits without necessarily suffering gold losses. When it started to lose gold in the 1960s, moreover, it negotiated ad hoc arrangements and agreements that encouraged other countries to hold dollars instead of buying gold; see Coombs (1976) and Solomon (1982). Accordingly, the United States was not obliged to deal quickly with its balance-of-payments problem. In the words of Charles de Gaulle, it enjoyed the 'exorbitant privilege' of using its domestic money to pay its foreign bills.

Second, the reserve-creating arrangements of the Bretton Woods System posed a basic threat to the viability of the system – a point made emphatically by Robert Triffin (1960). Because the IMF could not create international money – the Keynes plan had been rejected – the United States *had* to run balance-of-payments deficits to supply reserves to the rest of the world. As it did so, moreover, its net reserve position was likely to deteriorate; its dollar liabilities were apt to grow faster than its gold stock. Any such deterioration, moreover, was bound to impair the credibility of the US promise to sell gold for dollars, reduce the attractiveness of the dollar as a reserve asset, and wreck the reserve-creating arrangement on which the system depended.

Triffin's critique of the gold–dollar standard and his own plan for reform produced a torrent of other proposals (see, for example, Grubel 1963) and led eventually to a promising reform. In 1968, governments adopted the First Amendment to the Articles of Agreement of the IMF, allowing the Fund to create a new reserve asset, the Special Drawing Right (SDR), when and if this was required to meet the demand for reserves. The value of the SDR was defined initially in terms of gold (in a manner that priced it at one US dollar). In 1976, however, the Second Amendment to the Articles of Agreement took the IMF off gold by making the SDR the official standard of value, and the value of the SDR itself was redefined in terms of a basket of national currencies.

Small amounts of SDRs were actually created in 1970–1972 and 1979–1981. But the SDR arrived on the monetary scene too late to forestall the collapse of the Bretton Woods System, and has never acquired a major role in the international monetary system.

The Collapse of the System

In 1960, when Triffin published his attack on the gold-exchange standard, the US reserve position

was very strong; US gold holdings were far larger than US liabilities to foreign governments and central banks. But the balance-of-payments deficits of the 1960s eroded its reserve position, fulfilling Triffin's prophecy. The collapse of the Bretton Woods System, however, was not due to this development alone. It reflected the gradual deterioration in the competitive position of the United States, exacerbated by the economic consequences of the Vietnam War. By the late 1960s, the United States had ceased to be the stable centre of the monetary system; its inflation rate was rising, and its trade surplus was vanishing.

The first major break in the commitment to pegged exchange rates came in 1969. Rumours that the Deutschemark would be revalued vis–àvis the dollar attracted huge amounts of speculative capital to Germany and caused the German authorities to let the Deutschemark float rather than accumulate more reserves and thus increase the German money supply. The Deutschemark appreciated by 10% during the next 4 weeks, after which the German authorities converted the appreciation into a revaluation by pegging the Deutschemark–dollar rate close to its new market level.

The fatal break came in 1971, when the US payments deficit widened suddenly. It ran at an annual rate of $20 billion during the first quarter of 1971, four times as large as it had been in any previous calendar year, producing new rumours that the Deutschemark would be revalued. On a single day in May, the German authorities had to buy more than $1 billion in the foreign-exchange market to keep the dollar from depreciating, and they had to buy a similar amount during the first hour of the next day's trading. Therefore, they quit and permitted the Deutschemark to float again.

An appreciation of the Deutschemark, however, could not solve the basic problem – the very large increase in the US payments deficit – and American officials began to look for the best way to achieve a general exchange rate realignment. They did not want to raise the dollar price of gold, the only option open to them unilaterally. That would break faith with the governments that had held dollars rather than gold, and it might not work. A higher dollar price for gold would not devalue the dollar in a meaningful way unless other governments agreed to raise the dollar prices of their currencies. (On the discussions within the US government, see Solomon 1982; Gowa 1983; Leeson 2003.)

The crisis came to a head in August, after France had bought gold from the United States to repay a drawing on the IMF, and there were rumours of a large gold purchase by the Bank of England. The rumours were inaccurate but influential. On 15 August 1971, President Richard Nixon announced major changes in US policies. He froze wages and prices temporarily to combat inflation and asked Congress to approve an investment tax credit to stimulate output and employment. He imposed a 10% tax on imports and instructed the Secretary of the Treasury to close the gold window – to suspend US purchases and sales of gold.

The last two measures were designed to achieve an exchange rate realignment. They imposed two penalties on any foreign government that refused to revalue its currency. Its exports would be penalized by the tariff, and it could no longer count on buying gold when it purchased dollars in the foreign-exchange market to keep its currency from appreciating. The United States was widely criticized for adopting 'shock tactics' and breaking the rules of the trading system as well as those of the monetary system. But the tactics worked. In the weeks following the President's speech, several governments joined Germany in letting their currencies float temporarily, and after 3 months' bargaining a meeting at the Smithsonian Institution in Washington agreed to realign exchange rates formally. Most of the major industrial countries revalued their currencies against the dollar, and the United States devalued the dollar against gold. (It did not reopen the gold window, however, so that the new official price of gold was purely notional – the one at which the US Treasury would *not* buy or sell.)

The new pegged-rate regime, however, fell apart rapidly. The pound sterling was allowed to float in June 1972, and the end of the Bretton Woods System came early in 1973, after an

attempt by the United States to negotiate a second exchange- rate realignment. Japan allowed the yen to float in February, and six members of the European Community agreed in March to allow their currencies to float jointly.

These measures were seen to be temporary at the time, but governments soon came to believe that it would be impossible to return to pegged exchange rates, especially after the oil shock of 1973–1974 and the economic problems it produced. In 1976, the Second Amendment to the Articles of Agreement of the IMF replaced the original commitment to pegged exchange rates with much looser obligations. Governments would be free to choose any exchange rate arrangement *except* a fixed gold price, and the IMF was told to 'exercise firm surveillance over the exchange rate policies of members' (Articles of Agreement, Art. IV (3)) but was not told how to do that.

Although the term 'Bretton Woods System' is usually used to characterize the monetary system that prevailed until the early 1970s, a few have used it to describe a far more recent regime, which they describe as Bretton Woods II (Dooley et al. 2003, 2004). What do they mean? Throughout the 1960s, the United States ran balance-of-payments deficits because net capital outflows from the United States exceeded the US current-account surplus. In recent years, the United States has run balance-of-payments deficits because the US current-account deficit has exceeded net private capital inflows into the United States, and there has been as a result a huge accumulation of dollar reserves by countries that have been reluctant to let their currencies appreciate, most notably China, other East Asian countries, and the main oil-exporting countries. Many economists have warned that this payments pattern is unsustainable; see, for example, Obstfeld and Rogoff (2005) and Roubini and Setser (2004). The dissenters, however, compare it to the payments pattern of the late 1950s and early 1960s, which lasted for a decade before the Bretton Woods System collapsed. They maintain that the surplus countries, especially those in Asia, have chosen deliberately to hold down the dollar values of their currencies and thereby accumulate dollar reserves because they count on export growth to foster rapid output growth and thus the transformation of their national economies. There is, of course, no way to resolve this controversy. Time alone can do that.

See Also

- ▶ International Financial Institutions (IFIs)
- ▶ International Monetary Fund
- ▶ World Bank

Bibliography

Artis, M., and S. Ostry. 1986. *International economic policy coordination*. London: Royal Institute of International Affairs and Routledge and Kegan Paul.

Coombs, C. 1976. *The arena of international finance*. New York: Wiley.

Cooper, R. 1968. *The economics of interdependence*. New York: McGraw-Hill for the Council on Foreign Relations.

Cooper, R. 1972. Eurodollars, reserve dollars, and asymmetries in the international monetary system. *Journal of International Economics* 2: 325–344.

Dam, K. 1982. *The rules of the game*. Chicago: University of Chicago Press.

de Vries, M. 1976. *The International Monetary Fund, 1966–1971*. Washington, DC: IMF.

de Vries, M. 1986. *The International Monetary Fund, 1972–1978*. Washington, DC: IMF.

de Vries, M. 1987. *Balance of payments adjustment, 1945 to 1986: The IMF experience*. Washington, DC: IMF.

Dell, S. 1981. *On being grandmotherly: The evolution of IMF conditionality*, Essays in international finance, vol. 144. Princeton: International Finance Section, Princeton University.

Dooley, M., D. Folkerts-Landau, and P. Garber. 2003. *An essay on the revived Bretton Woods System*, Working Paper No. 9971. Cambridge, MA: NBER.

Dooley, M., D. Folkerts-Landau, and P. Garber. 2004. *Direct investment, rising real wages and the absorption of excess labour in the periphery*, Working Paper No. 10626. Cambridge, MA: NBER.

Eichengreen, B. 1991. *Golden fetters: The gold standard and the great depression, 1919–1939*. New York: Oxford University Press.

Eichengreen, B. 2006. *Global imbalances and the lessons of Bretton Woods*. Cambridge, MA: MIT Press.

Gardner, R. 1969. *Sterling–dollar diplomacy*. London: Oxford University Press.

Gowa, J. 1983. *Closing the gold window*. Ithaca: Cornell University Press.
Grubel, H. (ed.). 1963. *World monetary reform*. Stanford: Stanford University Press.
Horsefield, J. 1969. *The International Monetary Fund, 1945–1965*. Washington, DC: IMF.
International Monetary Fund. 1984. *Issues in the assessment of the exchange rates of industrial countries*, Occasional Paper No. 29. Washington, DC: IMF.
James, H. 1996. *International monetary cooperation since Bretton Woods*. Washington, DC/New York/London: IMF/Oxford University Press.
Kaplan, J., and G. Schleiminger. 1989. *The European Payments Union*. Oxford: Clarendon.
Kenen, P.B. 1986. *Financing, adjustment and the International Monetary Fund. Studies in international economics*. Washington, DC: Brookings Institution.
Kenen, P.B. 1989. *Exchange rates and policy coordination*. Manchester: Manchester University Press.
Keynes, J.M. 1925. The economic consequences of Mr. Churchill. In *The collected writings of John Maynard Keynes*, Essays in persuasion, vol. 9. London: Macmillan.
League of Nations. 1944. *International currency experience*. Geneva: League of Nations.
Leeson, R. 2003. *Ideology and the international economy: The decline and fall of Bretton Woods*. Basingstoke: Palgrave Macmillan.
Meade, J. 1951. *The balance of payments*. London: Oxford University Press.
Mundell, R. 1969. Problems of the international monetary system. In *Monetary problems of the international economy*, ed. R. Mundell and A. Swoboda. Chicago: University of Chicago Press.
Nurkse, R. 1945. *Conditions of international monetary equilibrium*, Essays in international finance, vol. 4. Princeton: International Finance Section, Princeton University.
Obstfeld, M., and K. Rogoff. 2005. Global current account imbalances and exchange rate adjustments. *Brookings Papers on Economic Activity* 2005(1): 67–123.
Roubini, N., and B. Setser. 2004. *The U.S. as a net debtor: The sustainability of the US external balances*. New York: Mimeo, School of Business, New York University, and University College Oxford.
Solomon, R. 1982. *The international monetary system, 1945–1981*. New York: Harper and Row.
Triffin, R. 1957. *Europe and the money muddle*. New Haven: Yale University Press.
Triffin, R. 1960. *Gold and the dollar crisis*. New Haven: Yale University Press.
Whitman, M. 1974. The current and future role of the dollar: How much symmetry? *Brookings Papers on Economic Activity* 1974(3): 539–583.
Williamson, J. 1983a. *The exchange rate system*, Policy analyses in international economics, vol. 5. Washington, DC: Institute for International Economics.
Williamson, J. (ed.). 1983b. *IMF conditionality*. Washington, DC: Institute for International Economics.

Bribery

Susan Rose-Ackerman

Abstract

Bribery is a form of rent-seeking meant to induce officials to serve private interests. Principal–agent relations are at the heart of the economic analysis of the subject. Bribery undermines government functioning by influencing electoral outcomes, lowering the benefits from public contracts, distorting the allocation of public benefits and costs, and introducing delay and red tape. Empirical work documents the negative consequences of corruption, and economic theory helps one understand the underlying incentives for payoffs.

Keywords

Bribery; Bureaucracy; Campaign finance, economics of; Corruption; Democracy; Presidential democracy; Principal and agent; Privatization; Procurement, government; Proportional representation; Public interest; Public services; Rent seeking

JEL Classifications

H8

Bribery and corruption are a form of rent seeking meant to induce official agents to serve the interests of those making payoffs.

Principal–agent relations are at the heart of the economic analysis of bribery. Payoffs induce agents to go against the interests of their principals, be they higher-level officials, politicians, or the citizenry in general. Bribery undermines the interests of principals by influencing electoral outcomes, lowering the benefits from public contracts, distorting the allocation of public benefits and costs, and introducing delay and red tape. The study of bribery thus highlights the conflict between the public interest and the market.

Widespread bribery can transform government actions ostensibly based on democratic or meritocratic principles into ones based on willingness-to-pay.

The theory of perfect competition emphasizes the impersonality of all market dealings. A manufacturer will sell to all customers irrespective of their race, gender, or inherent charm. Similarly, the ideal official makes decisions on the basis of objective, meritocratic criteria and is not influenced by personal, ethnic or family ties. Bribes can replace an impersonal meritocratic procedure with an impersonal willingness-to-pay procedure, or payoffs can support a system of personalized favours based on close personal relations. Alternatively, bribery can replace a personalized system based on family and ethnic ties with one based on financial capacity.

Early economic work on bribes concentrated on their role as prices and argued that they enhanced the efficiency of government (Leff 1964). This perspective has been overtaken by both theoretical and empirical work arguing for and documenting the costs of systemic corruption. On the theory see, for example, Rose-Ackerman (1978), Shleifer and Vishny (1993), and the literature reviewed in Bardhan (1997) and Rose-Ackerman (1999). Cross-country empirical studies are reviewed in Graf Lambsdorff (2006) and Rose-Ackerman (2004, pp. 303–10). Kaufmann and Kraay (2002), part of a World Bank Institute governance team, deal with the issue of whether high corruption causes low growth or whether low growth generates corruption. They conclude that the causal arrow runs from high corruption to low growth, but the issue remains vexed and has led to a turn to history to seek independent causes. The problem with econometric studies that use historical data, however, is that they cannot be a guide to policy. If one is concerned with reform, it seems necessary to engage with the messy real world of feedback loops and multiple causes. History can then be put to different use as a source of case studies of successful and failed reform efforts (Glaeser and Goldin 2006).

Corruption arises under many conditions in modern states. This article considers three variants: political corruption, kickbacks in major procurement and privatization contracts, and corruption in the allocation of benefits and burdens (for more details and references to the literature see Rose-Ackerman 1978, 1999, 2004, 2006).

Political Corruption

Non-democratic states tend to be more corrupt than democratic states, but democracies are clearly not immune from corruption. Obviously, corruption that arises from the competition for public office will be more prominent in democracies. The empirical results suggest that it is only long-established democracies that are less corrupt than other systems. As an example, the transition from socialism to market democracy in eastern Europe and central Asia has been fraught with corruption. During the transition, payoffs were a way to deal with an uncertain and rapidly changing environment just as, in the past, they had been a response to the excessive rigidities of a planned economy.

Furthermore, even within the universe of democracies, corruption levels vary with the constitutional structure of government. Kunicová and Rose-Ackerman (2005) find that presidential systems with legislatures selected by proportional representation are more subject to corruption than other democratic forms. Their explanation for this phenomenon is a bargaining situation in which a few strong party leaders negotiate with a powerful chief executive to share the spoils of office subject to relatively ineffective checks from voters, minority parties, and rank-and- file legislators.

At the individual level, the corruption of elected politicians depends upon the trade-off between their desire for re-election and their interest in monetary gain. Suppose voters are well-informed about politicians' votes but cannot observe bribes directly. Assume that politicians run for re-election on their voting record and that no campaign spending is needed. Then a bribe designed to change a vote in the legislature will cost the politician some constituency support. Bribes must be sufficient to compensate for the reduced chance of re-election. *Ceteris paribus,*

politicians with the lowest reservation bribes are those who are either quite certain of being elected or quite sure of defeat; in each case a decline in electoral support has little impact on the ultimate outcome. The closer the race, the higher will be the politician's reservation bribe.

In this simple model there is no need for campaign contributions, so bribes are used only for personal gain, and there is a direct trade-off between bribes and the probability of re-election. If payoffs *can* be used either to support a re-election campaign or as personal income, then all politicians may be corruptible, depending on their moral scruples and the salience of the issues influenced by corruption (Rose-Ackerman 1978, pp. 15–58). In electoral democracies, the control of corruption requires that re-election-seeking politicians feel insecure about their prospects but not too insecure. Too much security of tenure furthers corrupt arrangements. Too much insecurity can have the same effect.

Procurement and Privatization

No bribes occur in a perfectly competitive market, where suppliers can sell and demanders can buy all they wish at the going price. If bribes are offered, there must be some prospective excess profits out of which to pay them, and, if bribes are accepted, it must be because the agent's superiors are either privy to the deal themselves or else cannot adequately monitor the agent's behaviour. Corruption requires market imperfections. These are widespread in government procurement, resource concessions, and the privatization of public firms. The government will often be a monopsony purchaser or a monopoly seller; and it may need products not available 'off the shelf' so that a negotiated contract is necessary.

One might argue that corruption in procurement and the sale of assets furthers efficiency because the most efficient firm will have the highest prospective profits and so be willing to pay the highest bribe. This is simplistic. First, a winning firm in a procurement contract may gain advantage by lowering quality in subtle ways, not immediately obvious to government inspectors. Second, if managers of firms differ in respect for the law, the most unscrupulous have an advantage. Third, keeping payoffs secret both wastes resources and causes the market to operate poorly because of the low level of available information. Finally, the desire for payoffs may induce officials to contract for overly costly one-of-a-kind projects capable of hiding large kickbacks and to privatize firms on terms that favour corrupt bidders.

Mandating more effective competition is not always an option. In such situations one must consider the role of detection and punishment. Becker and Stigler (1974) first applied work on the economics of crime to corrupt payments. They stress the importance of giving each employee a stake in his or her job by, for example, providing non-vesting pensions. This will make workers less likely to take risks that could lead to their dismissal. More generally, the expected punishment for bribery should be tied to the marginal gain from marginal increases in the payoff (Rose-Ackerman 1978, pp. 109–35, 1999, pp. 52–9). Otherwise only some bribes will be deterred. Thus the marginal expected penalty for the bribe-taker, that is, the probability of apprehension and conviction times the penalty if convicted, must rise by at least 1 dollar for every dollar increase in expected payoff. If it does not, then even if a large lump-sum penalty is levied, only relatively small bribes may be prevented. The bribe-payer's marginal penalty should be tied, not to the size of the bribe, but to the marginal increase in profit that a bribe makes possible. Penalties set at a multiple of the bribe paid may have little deterrent effect on bribe-payers if the expected profits are many times larger.

Dispensers of Benefits and Burdens

Low-level officials frequently have considerable discretion to decide who should receive a scarce benefit such as a unit of public housing, expedited access to an important person, a liquor licence, or assignment to a particular judge. Others, such as health and safety inspectors, tax collectors, and the police, have the power to impose costs and the discretion to refuse to exercise that power.

Although legal pricing systems can sometimes substitute for payoffs here, in many cases there is a strong public policy reason for opposing a market solution.

How then can corruption be controlled? There are many ways to limit the discretion of officials to extract payoffs (Rose-Ackerman 1999, pp. 39–68). Consider just one option: the introduction of competitive pressures (Rose-Ackerman 1978, pp. 137–66). If a bureaucracy dispenses a scarce benefit, competition can be introduced by permitting an applicant to reapply if he has been turned down by one official. Then if the cost of reapplication is small, the first official cannot demand a large bribe in return for approving the application; in fact the offered bribe may be forced down so low that the official may turn it down and instead behave honestly. A few honest officials in this system may produce honesty in the others. Notice, however, that unqualified applicants will still wish to make payoffs, and their willingness-to-pay increases if they expect that most other officials to whom they could apply are honest.

The case for competition among inspectors or police is somewhat different and depends upon the feasibility and cost of overlapping authority. Thus, the operator of a gambling parlour will not pay much to a corrupt policeman if a second independent policeman is expected to come along shortly. The whole precinct must be on the take, that is, monopolized, to make high bribes worthwhile.

In short, the role of competitive pressures in preventing corruption may be an important aspect of a strategy to deter the bribery of low-level officials, but it requires a broad-based exploration of the impact of both organizational and market structure on the incentives for corruption facing both bureaucrats and their clients.

See Also

- ▶ Directly Unproductive Profit-Seeking (DUP) Activities
- ▶ Political Institutions, Economic Approaches to
- ▶ Principal and Agent (i)
- ▶ Principal and Agent (ii)
- ▶ Rent Seeking

Bibliography

Bardhan, P. 1997. Corruption and development: A review of issues. *Journal of Economic Literature* 35: 1320–1346.

Becker, G.S., and G.J. Stigler. 1974. Law enforcement, malfeasance, and compensation of enforcers. *Journal of Legal Studies* 3: 1–18.

Glaeser, E.L., and C. Goldin. 2006. *Corruption and reform: Lessons from America's economic history*. Chicago: Chicago University Press.

Graf Lambsdorff, J. 2006. Causes and consequences of corruption: What do we know from a cross-section of countries? In Rose-Ackerman (2006).

Heidenheimer, A.J., M. Johnston, and V.T. LeVine, eds. 1980. *Political corruption: A handbook*. New Brunswick: Transaction Publishers.

Kaufmann, D., and A. Kraay. 2002. Growth without governance. *Economica* 3: 169–229.

Kunicová, J., and S. Rose-Ackerman. 2005. Electoral rules and constitutional structures as constraints on corruption. *British Journal of Political Science* 35: 573–606.

Leff, N. 1964. Economic development through bureaucratic corruption. *American Behavioral Scientist* 8 (3): 8–14.

Rose-Ackerman, S. 1978. *Corruption: A study in political economy*. New York: Academic.

Rose-Ackerman, S. 1999. *Corruption and government: Causes consequences and reform*. Cambridge, UK: Cambridge University Press.

Rose-Ackerman, S. 2004. Governance and corruption. In *Global crises, global solutions*, ed. B. Lomborg. Cambridge: Cambridge University Press.

Rose-Ackerman, S., ed. 2006. *International handbook on the economics of corruption*. Cheltenham: Edward Elgar.

Shleifer, A., and R. Vishny. 1993. Corruption. *Quarterly Journal of Economics* 108: 599–617.

Brick, Laurits Vilhelm (1871–1933)

Hans Brems

Birck was born on 17 February 1871 in Copenhagen. He took his degree in economics at the University of Copenhagen in 1893, travelled in the United States in 1893 and in Britain and France in 1898–9. He served as a member of parliament and was active in wartime price control and postwar royal commissions on financial collapse and the

great depression. He taught economics and public finance at his alma mater 1903–33 and died on 4 February 1933 in Copenhagen.

Birck received the foundations of his theory of value (1902, 1922) from his teacher Harald Westergaard, who in turn had received them from Jevons. Jevonian households were do-it-yourself households engaged in barter, and to Birck their positive and negative utilities remained cardinal to the end. To Jevons Birck added Marshall, whom he considered the greatest name in our discipline. Marshall separated industries from households. By keeping his firms and industries small, he could justify a *ceteris paribus* assumption and consider the supply and demand curves of a competitive industry to be independent of the rest of the economy, and hence of each other. The curves would intersect in two-dimensional, simple and tidy partial equilibria. Birck applied such equilibria to case studies in 1909 and 1915 of 12 important commodities: coffee, flour, grain, kerosene, matches, meat, potash, potatoes, powder, salt, sugar and tobacco. Applied to statistical and historical data, theory – however simple – came to life, and Birck was at his best.

Birck's theoretical method, the numerical example, was exemplified by his massive *Virksomhed* (1927–8) and, in English, by his variation (1927) on Wicksell's theme that the capitalist saver is the friend of labour, though the technical inventor is not infrequently its enemy.

Selected Works

1902. *Værditeori: En Analyse of Begrebet Efterspørgsel og Tilbud* (Value theory: An analysis of the concept of demand and supply). Copenhagen: Søtofte.

1909. *Sukkerets Historie: En handels- og finanspolitisk Studie* (A history of sugar: A study in trade and public finance). Copenhagen: Gad.

1915. *Vigtige Varer: Deres Fremstilling, Forhandling og Beskatning* (Important commodities: Their production, distribution and taxation). Copenhagen: Børsens Forlag.

1922. *The theory of marginal value*. London: Routledge; New York: Dutton.

1927. Theories of over-production. *Economic Journal* 37: 19–32.

1927–8. *Den Økonomiske Virksomhed* (Economic activity). Copenhagen: Gad.

Bright, John (1811–89)

William D. Grampp

Keywords

Cobden, R.; Corn Laws; Female suffrage; Manchester School; Mill, J. S.

JEL Classifications

B31

John Bright, a Lancashire mill-owner, became a national figure in the campaign that repealed the Corn Laws in 1846 and that came to be known as the Manchester School.

Elected to the House of Commons in 1843, he continued to represent industrial constituencies most of his life and worked tirelessly for radical reform which to him meant reducing the scope of government, making it more representative and keeping its foreign policy peaceful. He was a man of strong views but not doctrinaire or unwilling to change them.

Believing in the market, he opposed factory legislation but not as it applied to children. At one time he supported John Stuart Mill's effort to give women the vote but later opposed the idea. He was against a state church, yet proposed its funds be distributed to all denominations as a once-and-never-again subsidy which recalls Smith's artful scheme. Although a Quaker, he never condemned war in principle and said that violence, while rarely called for, was sometimes necessary.

In his day Bright was said to be the pacifist who could have been a pugilist if he had not been a

Quaker. He does evoke truculence but what stands out a century later is his honesty and fierce independence. He combined them with an extraordinary speaking ability – in turns eloquent, persuasive, charming, brutally frank, cogent, and clever – all of which he could be because he had a first-rate mind. Never quite the equal of his intimate friend and ally, Richard Cobden, he nevertheless was one of the great figures in the reform movements of the century.

See Also

► Manchester School

Selected Works

1866. *Speeches on political reform.* London: Simpkin.
1878. *Speeches on questions of public policy,* ed. J.E. Thorold Rogers. London: Macmillan.
1879. *Public addresses of John Bright,* ed. J.E. Thorold Rogers. London: Macmillan.

Britain, Economics in (20th Century)

K. Tribe

Abstract

The foundations of the modern discipline of economics were Marshallian, symbolized by acknowledgement of his *Principles of Economics* (1890) as the leading English-language exposition, and the creation of the very first undergraduate teaching course in economics in Cambridge in 1903. The work of Maynard Keynes made a similar, lasting international impression, although in the second half of the century a neoclassical synthesis became internationally predominant, a tendency that fostered an interest in technique rather than economic problems that would have been quite alien to Marshall and Keynes.

Keywords

Ashley, W. J.; Boulding, K. E.; Britain, economics in (20th century); Cannan, E.; Chamberlin, E. H.; Chapman, S.; Coase, R. H.; Devons, E.; Dobb, M. H.; Economics, teaching of; Edgeworth, F. Y.; Fawcett, H.; Flux, A.; Free trade; Gorman, T.; Granger, C. W. J.; Hansen, A.; Harrod, R. F.; Hayek, F. A.; Henderson, H.; Hewins, W. A. S.; Hutchison, T.; Innovation; Institute for Fiscal Studies (UK); Invention; Jewkes, J.; Kahn, R. F.; Keynes, J. M.; Keynesianism; Klein, L. R.; Knight, F. H.; Lipsey, R.; Marschak, J.; Marshall, A.; Meade, J. E.; Meredith, H.; Monopoly; Neoliberalism; Neo-Ricardians; Perfect competition; Phelps Brown, E. H.; Pigou, A. C.; Political Economy Club; Price, L.L.; Robbins, L. C.; Robertson, D.H.; Robinson, E. A. G.; Robinson, J. V.; Royal Economic Society; Samuelson. P. A.; Schneider, E.; Shackle, G. L. S.; Sraffa, P.; Stackelberg, H. von; Stone, J. R. N.; Walters, A.; Young, A. A

JEL Classifications

B2

During the early 1900s, economics in Britain completed its transformation from a science accessible to a literate public to an academic discipline that required specific training; to be a student of economics henceforth implied that one was a college or university student. The literature of economics matched this transition. It moved out of the sphere of public argument into the closed world of an increasingly specialized academic discipline. Although there was never a perfect match between the general development of economic thinking and the pool of thinkers, these thinkers were henceforth overwhelmingly employees of universities, paid to teach and think about modern economics. Consequently, the story of British economics in the 20th century is closely related to the advance of university institutions, and

within these institutions, the formation of new departments of economics. Well into the 1960s, universities, colleges and schools remained the principal employers of 'trained economists', for there were very few alternative openings for 'economists' in business or public administration. In turn, the extension of opportunities for British university economists to develop their interest in the subject was for most of the 20th century conditional upon their ability to recruit undergraduate students; for taught graduate programmes were likewise a feature of the last third of the century.

In the 1990s, with the reclassification of virtually all higher education as university education and the general deterioration of student–staff ratios, the relationship between teaching and research that had prevailed through the greater part of the century broke down. Given the late appearance of graduate programmes, 'teaching' had meant lectures and classes to undergraduates, shared between the staff; while from the 1950s to the 1980s a 'class' was no more than a dozen students, in Oxford and Cambridge individual supervision being the norm. It was also usual for the more senior members of the department to present the more elementary lectures, but they, like their junior colleagues, pursued research projects alongside their other duties, supplemented by spells of departmental research leave. This arrangement did not survive into the 1990s. Those economists seeking to pursue a research career (and hence retain their reputation as economists) required a succession of external research grants to sustain any ambition of career development; they sometimes no longer taught at undergraduate level at all. The incentive to deploy senior economists in undergraduate teaching, and hence stimulate an interest in the subject among a younger generation, was seriously compromised. Meanwhile, employers specifically interested in economics graduates usually only required a first degree of their recruits. A Master's qualification was overqualification for anything other than appointment to a technical economic job, while an economics Ph.D. was serious overqualification for anything other than university employment. Given the unattractiveness of university employment to gifted young people, the number of British students studying at this level slumped. This evolutionary development in university institutions coincided with an unrelated transition in the discipline, from a focus on economic problems to an emphasis upon the elaboration of technique. In Britain, as elsewhere, mainstream training in economics had become instruction in a set of mathematical or statistical techniques that might, or might not, illuminate the kind of economic issues with which a wider public outside the university was concerned. Early in the century economics had been propelled into British universities by widespread belief in its public purpose and utility. By the end of the century, the discipline had become dominated by technicians for whom such beliefs were less important. As we shall see, this evolutionary progression was also related to the post-war internationalization of economics, so that by the end of the century the idea of a specifically 'British' economics had become an empty one.

Systematic tuition in economic principles originated in Britain. The first three-year university course was the Cambridge tripos, founded in 1903. The London BSc (Econ.), centred on the newly formed London School of Economics, had preceded this in 1901, but was structured in such a way that specialization in economics was only one of a number of social science options; and economics was taught only during the first year, at a very elementary level, in the commerce degree initiated by Ashley in Birmingham in 1902. The Oxford PPE, linking the study of Philosophy, Politics and Economics, and in this particular order because it had first been proposed by philosophers and opposed by economists, was initiated in 1920 (Chester 1986, 34 ff.). Ultimately, the London degree had the greatest influence in advancing the study of modern economics – not simply because of the success of the LSE in attracting both students and funding, but because the external London degree offered students resident outside London, and in the wider Empire, the opportunity of studying economics. The new University Colleges of Leicester, Nottingham, Exeter, Southampton, Reading, Hull and Bristol offered

to their students of economics the external London BSc (Econ.); and a succession of London Professors, from Cannan through Benham, Stonier and Hague to Lipsey, wrote popular undergraduate textbooks which remained widely used until late in the century.

Alfred Marshall, arguing for his new Tripos, had appealed to the growing need of business and public administration for young recruits conversant with the new science; a plausible enough argument, but one that in practice took many years to realise (Groenewegen 1995, pp. 556–7). William Ashley, generally unenthused by modern economics, sought a parallel development with his Birmingham commerce degree, intended to place appropriately trained recruits in the middle levels of management. The ambitions of both men were thwarted by a general lack of interest on the part of British business and public administration in 'new men'. Business remained dominated by small- and medium-size family firms until the interwar years at the very least, and here a professional training in law or accountancy remained a more useful general qualification than a degree in economics or commerce. In the mid-1930s having a first class degree in economics from the University of Cambridge led nowhere in particular: Terence Hutchison, appointed in the 1950s to Birmingham's chair, worked as a *Lektor* at the University of Bonn before the war; Alexander Henderson, later Professor of Economic Theory at Manchester, took a year out but then replaced Kenneth Boulding as Assistant Lecturer in Edinburgh. Economics had become a university discipline, but a degree in economics was a qualification that had little cash value outside academia. Only with the general expansion of the university system in the 1950s did it become customary for bright undergraduates to become in turn graduate students and then junior members of staff – the path taken by Clive Granger at Nottingham, for example. This pattern of training and recruitment altered little until the 1970s when demand for trained economists on the part of financial institutions and public administration began to develop.

The Institutions – Cambridge, Oxford, LSE and the Provinces

The Cambridge Tripos was the first honours economics programme in the world because it was a key ambition of Alfred Marshall to establish the subject as a modern independent discipline, and he was in a position to realize this ambition. Appointed to the Cambridge Chair in 1884 in succession to Henry Fawcett, author of the Millian *Manual of Political Economy* (1863), Marshall published *Principles of Economics* in 1890, and in 1892 *Elements of Economics of Industry,* an abridged version of the *Principles* for use by students which proved extremely popular. Later in 1891, Marshall oversaw the founding of the British Economic Association (from 1902 the Royal Economic Society, RES) as a vehicle for the publication of the *Economic Journal (EJ),* the first number of which appeared in March 1891 (Tribe 2001). In the United States, the *Quarterly Journal of Economics* had been founded in 1887 as the house journal of Harvard economists, while the *Journal of Political Economy,* founded in 1892, would be a house journal for Chicago economists. Marshall believed that the broad reception of new economics in Britain required a publication 'open to all schools and parties', and not therefore tied to any one institution. Following the publication of his textbook as a foundation for teaching, the *EJ* provided a platform for discussion among economic specialists while also keeping them informed of new publications, the current contents of foreign journals, and other relevant developments. The Tripos was the third of Marshall's stones in the new edifice.

Principles and *Elements* were a runaway success in the English-speaking world. The *EJ* in its early years indeed published a wide range of economic opinion – including, for example, the Erfurt Programme of the German Social Democratic Party, Vol. I September, 1891, pp. 531–3. But the Tripos remained merely a pedagogic monument for many years: during the 1930s, as many as 60 per cent of those taking the one-year Part I achieved modest Thirds (Tribe 2000). Nonetheless, there were, during the 1930s, many graduates

whose later reputation as economists began in Cambridge. In the 1880s and 1890s, economics had been taught as an option within the History and the Moral Sciences triposes at Cambridge; Marshall had made himself deeply unpopular among his colleagues with his persistence in seeking a separate existence for the teaching of economics, and having granted his wish in 1902 they proceeded to purge all economics from their own curricula. The tripos was certainly a model of a free-standing economics degree, but even in the boom years of the later 1940s the number of annual Firsts and Upper Seconds in Part II (the final examination) more or less matched the number of eminent economists in the faculty. The tripos, for the first 50 years of its existence, proved more successful in supporting the largest concentration of academic economists in Britain than teaching economics to receptive students.

On the other hand, many of Cambridge's economists turned to writing introductory textbooks under the auspices of the *Cambridge Economics Handbooks* series. The first of the handbooks was Hubert Henderson's *Supply and Demand,* published in 1921, followed by Dennis Robertson on money (1922), Maurice Dobb on wages (1928), and Austin Robinson on the structure of industry (1931) among many others. Maynard Keynes took over the series in the mid-1920s, and drafted a general introduction printed in all editions arguing that economics was a method, not a body of doctrine, 'an apparatus of the mind, a technique of thinking, which helps its possessor to draw correct conclusions'. Keynes was here reiterating his belief in the *organon* as the core of the Marshallian legacy, 'a machinery that we build up in our minds, a method, an organon of enquiry that can be turned to particular problems as they arise. . .' (Pigou 1925, pp. 86–7); to which Keynes added the republican principle that the purpose of the *Handbooks* was to expound the elements of economics 'in a lucid, accurate and illuminating way, so that the number of those who can begin to think for themselves may be increased. It is intended to convey to the ordinary reader and to the uninitiated student some conception of the general principles of thought which economists now apply to economic problems'. Published in the United States and widely circulated in the Empire, some of the handbooks were also translated, emphasising the general absence at this time of similar short works suitable for students of economics, as well as the manner in which Cambridge economics, generally unreceptive to the development of economic thinking elsewhere in Britain and abroad, was nonetheless projected into a wider world.

Oxford economics followed a different path. It had been the centre of British economics in the 1880s, pursuing the development of extension teaching in many provincial centres and graduating among others Edwin Cannan, W.J. Ashley, L.L. Price and W.A.S. Hewins (Kadish 1982, ch. 2). But Francis Edgeworth, appointed to the Drummond Chair in 1892, entirely lacked Marshall's institutional ambition, and in any case did not share Marshall's view that an understanding of economics required three years of systematic tuition. During the early 1900s teaching in Oxford remained broadly Millian (Young and Lee 1993, p. 7), with Marshall being reserved for the more advanced students. The background of those who taught was primarily in history – when Roy Harrod was elected fellow of Christ Church in 1922, it was to a fellowship in history, but he immediately took himself off to Cambridge to study with Keynes, and then on his return arranged for Edgeworth to provide informal graduate supervision. By the later 1920s, with student numbers growing, new appointments were predominantly PPE graduates, among them Henry Phelps-Brown and James Meade in 1930. John Hicks had graduated in 1926 from the PPE, but with a second-class degree and was very fortunate to get taken on at the London School of Economics (LSE), since that institution too was beginning to recruit staff from among the ranks of its own graduates. Oxford lacked the organizational thread that the tripos gave Cambridge economics, and had no central figure to match Keynes, but it was perhaps as a consequence more open to external developments. In 1935 Jacob Marschak, an Oxford lecturer since he had been stripped of his Heidelberg post in 1933, was appointed to a readership in statistics and was made founding Director of the Institute of Statistics. Although, the

institute was not the first of such research bodies established in Britain – Manchester's Research Section under John Jewkes preceded it – its foundation predated any plans for Cambridge's own Department of Applied Economics which, delayed by the war, eventually began work in 1945. Also significant is that fact that the Institute was funded externally, by the Rockefeller Foundation, together with a number of new posts in the social sciences. Similarly, Lord Nuffield's benefaction of the later 1930s – he had approached the university with the idea of funding a new engineering college and was persuaded by the then Vice-Chancellor, A.D. Lindsay, of the need for a social science foundation – also provided a focus for collaborative research in economics that Cambridge lacked. In 1941, the Nuffield College Committee established a social reconstruction survey, while the Institute conducted studies on full employment. This complemented work that had been initiated in the mid-1930s by the Oxford Economists' Research Group, again funded with Rockefeller money, which conducted studies of business decision-making and the role of interest rates, this work being published in the first issue of *Oxford Economic Papers* in 1938.

By this time, the *EJ* was being edited from Cambridge by Keynes and Austin Robinson and was widely, and disparagingly, referred to as the *Cambridge Economic Journal,* while the RES had also become closely associated with Cambridge. The LSE had also founded its own journal, *Economica*, in 1920, and with the launch of the 'new series' in 1933 this became a dedicated economics journal. This coincided with the maturation of a style of work distinct from Cambridge, by the mid-1930s condensed into a general scepticism of the significance of Keynes's *General Theory* and what today would be recognized as a strong leaning to neoliberalism. The School had been established in 1895 with a legacy linked to the Fabian Society (Kadish 1993, p. 230), the common denominator being Sidney Webb and his involvement with commercial education in London. Before the First World War its teaching staff had been predominantly part-time – Cannan, its first professor of economics, retained his part-time status until his retirement in 1926 – but teaching was reorganized during the 1920s, adding a commerce degree to the BSc (Econ.) and replacing part-time with permanent staff recruited from among its own students. Lionel Robbins, appointed to the chair of economics in 1929, and Arnold Plant, who became professor of commerce the following year, were both examples of this trend, Plant gaining a First in economics in 1923 having also been awarded a First in commerce the previous year (Plant read for the commerce degree as an external student alongside his full-time study of economics). The arrival of Friedrich von Hayek in 1931 as visiting professor confirmed the neoliberal profile that LSE economics assumed from the 1930s to the 1950s, but also the openness of the institution. Cannan's successor as professor had been the Harvard economist Allyn Young, and there was widespread dismay when his early death from pneumonia in 1929 terminated a direct connection to American economists that had been expected to endure for many years.

Likewise, LSE was more catholic in its teaching and reading materials than any other British institution of the time – Frank Knight's *Risk, Uncertainty and Profit* was used as a central text and re-issued in 1933 as No. 16 in the School's reprint series. As a first-year undergraduate in 1948, Bernard Corry recalled being first given sections of Samuelson's *Foundations* to work through, followed by Erich Schneider on the theory of production, and Pallander on location theory (Corry 1997, pp. 179–80). In a 1937 survey of the School's work, Plant and Robbins noted that Frank Taussig's *Principles of Economics* was a 'good modern manual' which, besides specialized sections on public finance, railways and social reorganization, covered much the same ground as the LSE course in economics. Marshall's *Principles* headed the list of works on general economics (Plant and Robbins 1937, pp. 67, 69). At least part of the differences between Cambridge and LSE economists during the 1930s can be traced to this contrast between an LSE aggressively open to the international development of economics, and a Cambridge which simply assumed that it was in the van of such development and did not therefore need to take account of

work elsewhere. Acknowledging her debts on the opening page of *The Economics of Imperfect Competition,* Joan Robinson referred exclusively to Cambridge colleagues – Marshall, Pigou, Sraffa, Kahn, Austin Robinson and Gerald Shove. She did note the contributions to competition theory of Erich Schneider and Heinrich von Stackelberg, but considered that 'their work is marred by the use of unnecessarily complicated mathematical analysis where simple geometrical methods would serve' (Robinson 1933, p. vii).

By the 1930s, Cambridge was graduating 50–60 students from its Part II every year, and well over 100 students left the LSE annually with a BSc (Econ.) containing an increasingly variable amount of economics. The new universities founded from the turn of the century – Birmingham, Manchester, Liverpool and Sheffield – made little direct headway in finding a constituency of students eager to learn the new economics, but they did find a ready market for teaching in commerce, which contained some economics. In most cases this teaching was quite practical, covering law, banking, economic geography, history and languages; and railway management was often an important component, given the size of the railway companies and the numbers of their employees. For many students approaching economics for the first time, it was taught as part of a vocational course that had the support of significant local employers. This was especially true in Scotland, where the four ancient universities – Glasgow, Edinburgh, St. Andrews and Aberdeen – were closer to the Continental European model, law and medicine being a part of the university. Chartered accountants in Scotland took university courses in elementary economics, highlighting a natural link between the professions and the university absent in England.

Ashley returned to Britain from Harvard's new chair in economic history to found Birmingham's Faculty of Commerce in 1902, but although this has become the single most well-known example of commerce teaching in Britain, it was atypical in many ways. Ashley had ambitions for commerce analogous to Marshall's for economics, seeking to educate future management leaders rather than the future line managers and college teachers turned out in Liverpool and Manchester. He established an advisory board with local business in a deliberate effort to recruit the sons of business families. But instead of drawing on the local business community for the teaching of accounts, commercial law and banking as Liverpool or Manchester had done for many years, Ashley made accounting a professorial position and in 1906 followed this with a chair in finance. These posts were not justified by the student numbers that he recruited. There were never more than 36 students registered for the commerce degree before 1914, and total registrations only averaged in the high fifties once the short-lived post-war boom had passed. Birmingham's later reputation was based not on its early commitment to commerce, but on the coincidence that Frank Hahn, Alan Walters and Terence Gorman all taught there in the mid-1950s. Birmingham, together with Nottingham, was the first British institution to make a significant effort to develop mathematical and statistical analysis in economics.

Manchester was another important centre: it was here that the first university-based research section was established under Jewkes in the early 1930s, and Manchester economists predominated among those recruited to government service during the Second World War. The Faculty of Commerce had been established by Sydney Chapman (a former student of Alfred Marshall) in the late 1903, building upon a solid foundation of teaching in political economy most recently developed by Alfred Flux, but reaching back to Jevons's classes in the 1870s. Degrees were offered in both commerce and honours economics, Chapman using part-time local professionals for the more specialized parts of the commercial curriculum and appointing young economists to do the non-specialized teaching. This strategy enabled him to develop the teaching of economics, and many of the pre-First World War junior staff went on to chair their own departments: Hugh Meredith taught in Manchester 1905–8, and then was professor at Queen's Belfast from 1911 to 1945; Robert Forrester taught in the Faculty 1910–13, went to Aberdeen, then the LSE, and was Professor at Aberystwyth from 1931 to 1951; Harold Hallsworth taught in Manchester during 1910,

later becoming Professor at Newcastle; Douglas Knoop taught in 1909, became a lecturer in Sheffield in 1910 and was then later Professor from 1920 to 1948; A.N. Shimmin taught 1913–15, and was from 1945 professor of social science at Leeds. Clearly, Manchester became an important staging post in the development of careers which imposed a clear pattern on the development of the teaching of economics in provincial Britain, and hence by extension the propagation of economic understanding to a diverse range of students.

This pattern in the academic life cycle had important consequences for the advancement of economics in 20th-century Britain. Those appointed to junior posts in this initial phase of pre-First World War expansion quickly moved on to more senior posts as new departments were established, but they then stayed in them for many years. This blocked mobility during the later 1920s and 1930s. But many senior members in this first cohort retired together in mid-century, creating an opening for renewal in the organization of academic economics, reinforced by increased demand for the teaching of economics in the late 1940s. During the immediate post-war period departments expanded to meet this demand; new posts were created, and a fresh wave of young candidates filled senior appointments. These in turn dominated university departments during the 1950s and early 1960s, but reached retirement age at about the same time that new universities were being founded and the number of senior positions extended once more. The pace of development of research and teaching in economics that took place in Britain during the 1960s rested to a considerable degree on the fluidity and openness that this academic life cycle created.

But these two successive surges – in the 1940s and the 1960s – of mobility, expansion and disciplinary development faltered with the uncertainties of the 1970s, and then broke on the university cutbacks of the 1980s. The mobility and advancement of younger staff trained in the later 1960s and early 1970s was blocked; this cohort grew old together in the same posts while bright young economists looked elsewhere for employment, and for which in any case they did not require to spend several years on a Ph.D. that an academic career now dictated. The average age of departments increased year by year, hollowing out the institutional hierarchy. By the 1990s, the pool of potential young British economists was severely depleted, given the small number of doctoral and postdoctoral students in the system; and with the slow resumption of recruitment the cycle simply skipped a generation expanded the pool from which it drew. Shortlists came to be dominated by applicants from the EU and beyond, attracted by the openness of the UK labour market and the experience of working in the English language. Graduate programmes likewise became dominated by foreign students. As with recruitment to medical staff in the National Health Service, British universities made good the manifest deficiencies of the British educational structure by turning for graduate students and faculty to those trained elsewhere.

The Interwar Years

The foregoing is not intended to substitute for a more orthodox 'history of economic thought' story. It instead demonstrates how the building of a discipline required a financial and institutional framework as a condition for the development of 'economic careers', which careers in turn provided the basis for the elaboration of economic argument as spoken, written and published discourse. The first movers in this latter process are indeed generally to be found in Oxbridge and London; but, for a discipline to flourish, followers are also needed, who in turn have access to a secure institutional structure. Hence, the importance of a national perspective upon the development of economics in Britain.

Cambridge did occupy centre stage in the first half of the century, partly as a consequence of the employment opportunities the new tripos presented: students had to be supervised and courses of lectures delivered, and this all added up to a significant number of college fellows and University lecturers. Marshall was also an important spiritual and pedagogic presence – after retirement in 1908, he continued his practice of open

hours at home for students, lending them the books that would later form the core of the Marshall Library. His young protégé Arthur Pigou had marked himself out early on with a number of articles in the *EJ* notable for their brevity and formal exposition – anticipations of a style that had not then become customary. His *Wealth and Welfare* broke new ground in seeking to determine what 'welfare' might be, and noting that however defined, if the 'National Dividend' (as he termed GNP) increased, then welfare also increased. Redistribution of welfare through the population could also be brought about, but given the regressive nature of the contemporary taxation system he thought of this chiefly in terms of access to health and education services. He noted that monopoly tended to distort the distribution of welfare, so that this book also involved an extended treatment of duopoly and imperfect markets. This and the work of Alfred Marshall had a considerable contemporary impact upon American discussion of price and competition, forming a natural background to the later work of Frank Knight and Edward Chamberlin, especially in respect of Pigou's observations on the level of equilibrium output under monopolistic competition (Pigou 1912, pp. 294, 356). The 1920 revision of this work into *Economics of Welfare* re-emphasized the social duties of the economist as outlined by Marshall in his inaugural lecture of 1887; and a new emphasis is laid upon the impact of taxation, commensurate with the consequences of the war for the post-war economy. The Marshallian cast of the work is highlighted by the following credo from the Preface:

> The complicated analyses which economists endeavour to carry through are not mere gymnastic. They are instruments for the bettering of human life. The misery and squalor that surrounds us, the dying fire of hope in many millions of European homes, the injurious luxury of some wealthy families, the terrible uncertainty overshadowing many families of the poor – these evils are too plain to be ignored. By the knowledge that our science seeks it is possible that they may be restrained. Out of the darkness light! To search for it is the task, to find it, perhaps, the prize, which the 'dismal science of Political Economy' offers to those who face its discipline. (Pigou 1920, p. vi)

Keynes certainly shared this credo, as his introductory comments to the Cambridge Handbooks show, but his later characterization of Pigou as a 'classical' that is, superseded, economist has subsequently been too easily subsequently accepted at face value. Pigou, being the professor, was debarred from supervising undergraduates, so that his involvement in teaching was limited to lecturing, and this he generally did at an elementary level only. As with many of his generation – D.H. MacGregor in Oxford, Alec Macfie in Glasgow – he had been badly affected by his experiences in the First World War, and played little further part in the shaping of teaching and research in Cambridge. He has consequently, and unjustly, been excluded from 'Cambridge view' of the history of economics, which has come to be dominated instead by Sraffa, Kahn and the Robinsons, amongst others (Collard 1981).

The *locus classicus* of this Cambridge 'insider story' is George Shackle's *The Years of High Theory,* although curiously Shackle was never a 'Cambridge man': he went to school there, but was never connected with the university. *The Years of High Theory* takes its departure from Sraffa's 1926 *EJ* article, and ascribes to contemporary non-Cambridge economists a dogmatic and universal belief in 'perfect competition'. Hence Sraffa's theoretical critique of perfect competition is presented as a radical, definitive, if unappreciated, settling of accounts, upon which new work can thereafter build. Here Shackle joins later neo-Ricardians, for whom likewise Sraffa is of decisive importance to the development of economic theory. 'Perfect competition' had however only just been systematically adumbrated, in Chapter 6 of Frank Knight' s *Risk, Uncertainty and Profit* (1921), and by no means dogmatically; indeed, Shackle imputes to British economists of the 1920s views more common in the America of the later 1940s, and not before.

Dennis Robertson also fails to register in the Cambridge story, despite having Keynes as his Cambridge Director of Studies, and then spending almost his entire working life in Cambridge, retiring in 1957. This neglect can be attributed to his later criticism of Keynes, describing in 1948 the *General Theory* as 'a step backwards' which prematurely embraced 'stagnationism' 'on the

strength of one bad depression' (Robertson 1948, p. xvi). Remarks such as these make his relative neglect all too understandable, but this should not be allowed to obscure the larger significance of his early work. Hitherto studies of economic cycles had focused on the periodicity of price movements (Morgan 1990, chs. 1, 2); the analysis of *Industrial Fluctuation* went behind price movements to the variations in output and employment that they represented. That bust follows boom was easily accepted; but why a slump should be followed by recovery was not so easy to explain. Robertson identified a number of causes, most important of which was invention and innovation, an emphasis which was new at the time in Britain, and which Robertson had arrived at without having read Joseph Schumpeter (Presley 1981, pp. 178–9).

Robertson's *Banking Policy and the Price Level* (1926) was likewise an influential work, extending his study of fluctuations to cover monetary phenomena (Laidler 1999, 93 ff.). Robertson's mannered writing style did not make this book any easier to read, but as Laidler points out, Pigou took over large sections of the argument in his own *Industrial Fluctuations* (1927), disseminating Robertson's ideas in more readable English. As with his first book, Robertson took his departure from observable facts – that the British banking system balanced deposit liabilities against short-term loans. The banking system was therefore charged with coordinating the public's short-term saving with firms requirements for working capital, and although he noted the forced saving involved in this, he also saw its potential as a stabilizing factor, moderate forced saving being therefore the price paid for progress.

Cambridge in the 1930s is however dominated by the figure of Keynes, and not only intellectually. He had resigned his University Lectureship in 1920, after which his formal connection to the university was solely as a college fellow. Nonetheless, he made up for Pigou's disengagement through his editorial work on the *EJ* with Austin Robinson, in the Political Economy Club, to which promising students were invited and required to ask questions of visiting speakers, through his work for the college, and through his engagement in the arts. In Cambridge lectures could be offered by any college fellow, and were not confined to faculty members. Keynes developed a practice of lecturing from the proofs of his next book, the experience obviously leading him to substantial revisions (Rymes 1989). He found jobs for some bright graduates – while other bright graduates of whom he was unaware found that their Cambridge First might not necessarily lead anywhere in particular (Tribe 1997, pp. 77, 129).

Keynes's reputation has long been overlaid with 'Keynesianisms' of various kinds. That his memorial service in 1946 was held in Westminster Abbey is indication enough that, whatever the nature of his reputation, it was a very great one. Much of his work in the 1920s took the form of superior economic journalism – from *The Economic Consequences of the Peace* (1919) that made his public reputation, through 'The Economic Consequences of Mr. Churchill' (1925) to 'Can Lloyd George Do It?' (1929). His rise to become the single most influential British economist of the century began in the early 1930s. Peter Clarke has provided a lucid account of the early part of this story: the nature of contemporary government policy, Keynes's evidence to the Macmillan Committee in 1930, its relation to the two volumes of the *Treatise on Money* published that year, the impact of the abandonment of the gold standard in September 1931 and of free trade over the winter of 1931–32, and the consequent genesis of a new *general* theory of employment, interest and money – there is little dispute about the main lines of these developments (Clarke 1988).

Argument breaks out however over the substance and intentions of the *General Theory,* published in February 1936. David Bensusan-Butt captures precisely the sense of confusion a modern reader experiences coming to this work for the first time:

> Never did a book fall more quickly and more completely into the hands of summarisers, simplifiers, boilers-down, pedagogues and propagandists. To get at what it seemed like at the time (and perhaps what it really was and is) one has to fight one's way through a cloud of commentators, and try to see it in a more empty landscape. (Quoted in Skidelsky 1992, p. 537.

Notoriously, Keynes was one of the earliest such commentators, reflecting on his intentions in an article in the *QJE* in February 1937. Although few would seriously dispute that the *General Theory* marks the inauguration of an integrated macroeconomics, it was built out of existing elements – and some at least of the disagreements engendered by the book can be related to incompleteness in the integration of these elements. David Laidler has also shown, for example, that one of the most general statements that can be made about the *General Theory* – that it provides a clear role for government not in substituting for market activity, but by influencing the expectations of investors and businessmen – adopts arguments already made in Lavington's *The English Capital Market* (1921) (Laidler 1999, pp. 87–8).

The translation of Keynes's fluent prose into the diagrams and algebra better suited to an increasingly formalized style of economic argument followed publication very rapidly. Brian Reddaway, reading a review copy of the book on the way to a post at Melbourne University arranged by Keynes, sketched four equations relating savings, income, investment, the rate of interest and the supply of money and published these in the June 1936 issue of *Economic Record* (Reddaway 1936). On 26 September 1936, at a meeting of the Econometric Society in Oxford, a session was devoted to the *General Theory.* Here Roy Harrod, James Meade and John Hicks made graphical and algebraic presentations, Hicks writing this up in his article 'Mr. Keynes and the Classics' published the following year (1937). Thus was born the classroom IS–LM presentation of Keynes's ideas (Young 1987).

The transformation of the *General Theory* into a blueprint for managing the mixed economy was, however, effected along two separate paths. In the United States Lawrence Klein, Alvin Hansen and finally Paul Samuelson systematized Keynes's insights and rendered them consistent with the new neoclassical economics (Klein 1948; Hansen 1953; Samuelson 1955). In Britain, the outbreak of war in 1939 and the entry of British economists, including Keynes, into government service provided a unique opportunity to deploy Keynes's insights in managing the wartime economy (Cairncross and Watts 1989, chs. 2–7).

The basic framework had been laid down by Keynes in his 'How to Pay for the War', reversing the assumptions upon which the *General Theory* had been built. The basic task now was to run an economy at its maximum potential output for war production without generating inflationary pressures. Such diverse characters as Lionel Robbins, Ronald Coase, Brian Reddaway, John Jewkes, Ely Devons and James Meade were recruited into government service to facilitate the wartime management of the UK economy. Whereas financing the First World War had been primarily a matter of managing international money markets – a task in which Keynes had played a part – 'paying for the war' now meant management of the domestic economy. Inflation was to be avoided as a means of suppressing private consumption in favour of war production. Excess purchasing power was instead to be absorbed through additional taxation, which implied estimation of the actual level of excess. A thorough system of rationing was devised, and financial planning increasingly gave way to manpower planning. Allowance had to be made for the subsidies necessary to stabilize the cost of living, and, on the assumption that this stabilized gross incomes, total volume of money demand needed to be established. By subtracting the amount of goods and services coming on the market an 'inflationary gap' could be identified, representing the amount of excess demand that had to be siphoned off. As early as the winter of 1940 government treated pressures in the economy in terms of an 'output gap' separating the level of demand from the capacity of factors of production to meet these demands (Sayers 1983, p. 106). The 1941 Budget broke new ground, presented in a national accounting framework that would enable such estimations to be made (Kaldor 1941, p. 181). Moreover, this approach implied that the primary economic aim of governments should be the stability and growth of national income, rather than the more narrowly financial considerations traditionally associated with reviews of government income and expenditure. This was underlined by the formulation of post-war plans such as William Beveridge's

Social Insurance and Allied Services (1942), followed by the *Employment Policy* White Paper of June 1944, the month of the Normandy landings (Coats 1993a, p. 558). It was this framework that wartime economists bequeathed to the peacetime civil servants who succeeded them, and which enabled them to manage the economy in terms of Keynesian aggregates. The Economic Section, the central body of economic advisors that had been led by Robbins for most of the war, survived the transition to peacetime, but with a much reduced role. Coats notes that fewer than 20 professional economists were employed by the government on matters relating to macroeconomic policy during the first two post-war decades (Coats 1993b, p. 523).

There have been many versions of Keynesianism since (Backhouse 2006), but the most misleading variant is that which links Keynes to the centralized management of peacetime mixed economies. Some sort of Keynesian consensus did prevail in the British academic establishment from the later 1940s until the early 1970s, but the overriding concern, which had brought its senior members into the discipline, was a belief that the depression of the 1930s should not be allowed to recur. 'Keynesianism' offered a route to a policy synthesis that could realize this, but this was not translated directly into the pursuit of 'Keynesian' economic policies on the part of post-war Labour and Conservative governments. The Economic Section was not ineffective in its advice, but it was very small; while academics outside Whitehall lacked direct influence on the formation and execution of policy, chiefly confined in their expression of opinion to the letters' column of *The Times*. Hugh Gaitskell had been an economics lecturer at University College London and published on capital theory in the *Zeitschrift für Nationalökonomie*, but the Labour Party was never in power during his period of leadership. Harold Wilson likewise came from an Oxford economics background; his incoming Labour Government of 1964 did establish a Department of Economic Affairs, but its chief task was the drafting of a National Plan on the French model. The drafting and execution of legislation right up to the early 1980s was conducted by generalist civil servants with no special background in economics, directed for the most part by Ministers likewise lacking in formal economic training. The 'Keynesian' nature of their approach to government and the economy derived not from any particular theoretical beliefs, but chiefly from a generalized public expectation that it was the job of government to counter downturns, stabilize employment and promote growth. Until 1979, any party that denied its capacity to fulfil such electoral expectations stood no chance of gaining office. Harold Wilson observed acutely that 'Whichever party is in office, the Treasury is in power', but there is now an extensive literature which documents the essentially pragmatic, rather than dogmatic, nature of Treasury decision-making during the 1950s and 1960s, supposedly the heyday of Keynesianism (Peden 1988).

The Post-war Legacy

During the 1930s a number of British economists made theoretical innovations of lasting significance. This was indeed the 'decade of high theory', to borrow from George Shackle, but it was certainly not, as he suggests in his book, an exclusively Cambridge preserve. Ronald Coase, who graduated with a commerce degree from LSE in 1932, went that same year to his first appointment in Dundee, where he drafted his essay identifying a firm as a replacement for market transactions, eventually published in 1937. John Hicks, having published in 1934 an article in which consumer preferences displaced utility, went on in *Value and Capital* (1939) to create a neoclassical microeconomic synthesis. James Meade published in 1936 his *Introduction to Economic Analysis and Policy,* the first of many seminal works. All later gained the Swedish Riksbank Prize in Economic Sciences (in 1991, 1972 and 1977, respectively) for these and other works. But what is most notable about these annual awards, made since 1969 and beginning with Ragnar Frisch and Jan Tinbergen, is that they are dominated by American economists who began their careers in the 1940s and 1950s. For in this period American economics became international economics.

The war itself had turned out to be the apotheosis of British economics. US foreign policy sought to block any prospect that post-war Britain would resume its former world role, and assumed Britain's former international stance as model democracy and proponent of free trade and economic liberty. Teaching of economics in American universities expanded, and during the 1950s graduate programmes were developed on this foundation. There was a parallel expansion in demand for courses in undergraduate economics in Britain, but neither the will nor the money to develop graduate education. Increasingly, bright students and young economists looked to American connections to develop their careers. Coase was already there; Alexander Henderson went from Manchester to Carnegie Mellon in 1950, and became joint author of the first textbook on linear programming; Clive Granger had by the early 1970s gravitated to California. In turn, the teaching of economics in Britain became increasingly modelled upon American programmes, increasingly making use of American books and articles (Backhouse 1996, 2000).

As already noted, with the end of the war the majority of economists had quickly left government employment and moved back into the university. Economics was widely regarded as a 'modern' subject in school and university (Coats 1993c); educational opportunity was widely understood as the path to social mobility, a belief underwritten by Lionel Robbins's report to the government which argued that extension of university access would not compromise entry standards or teaching (Committee on Higher Education 1963). This finding coincided with the opening of a number of new universities in which social sciences played a significant role. In 1964, Richard Lipsey moved from the chair at LSE to the founding chair at Essex, primarily because he saw the opportunity to develop the graduate economics programmes there that his colleagues at LSE had declined (Tribe 1997, 217 ff.). Once established, this model rapidly spread, but then ran into the uncertainties of the 1970s. As economics became more technical, the capacity to train students in the new techniques remained very restricted. Generational succession, as outlined above, also played a role as a new generation, born into the certainties of the 1950s and 1960s, found themselves in an uncertain world.

As Roger Middleton has argued, financial pressure on universities in the later 1970s and 1980s was coupled with a collapse in the public authority of universities (Middleton 1998, p. 312). Moreover, throughout the 1980s academic economists were, with a few notable exceptions, generally hostile to government policy. Notoriously, this was expressed in a letter to *The Times* in March 1981 where 364 economists signed up to the argument that government policy would deepen the current depression and slow recovery. This polarized politicians and economists, to the lasting cost of the latter (Backhouse 2000, p. 31). University economists were consequently shut out of government decision-making while at the same time a broader public found the increasingly technical preoccupations of economists of little relevance to an understanding of economic problems. The broad consensus that had in the 1950s and 1960s made economics the 'modern' discipline broke upon widespread popular disillusion with both modern economics and the universities within which it was practised.

The evolutionary development of the discipline was exacerbated by the process of research audit that began in the mid-1980s, ranking departments and their staff on the basis of research publications (the Research Assessment Exercise, RAE). Although this provides for a system of peer review and is not imposed by a separate educational bureaucracy, the resultant ranking was increasingly employed to determine the allocation of resources between and within universities. Furthermore, peer review has tended to sharpen the 'scientization' and public isolation of British economics, since 'professional' prestige and a high ranking comes only from publication in a very restricted number of international journals, not from an interest either in undergraduate education or in public issues (Middleton 1998, 221 ff.). Each subject area draws up its own schedule of approved publication media, and in the case of economics this list has always been weighted towards 'rigour', which was what economists

had come to pride themselves on as compared to the other social sciences. Since these other social sciences were less 'rigorous' in their judgement of what counted as worthwhile research outputs, median economics departments assessed in the 2001 RAE fared very badly within social science faculties, losing funding and strengthening the polarizing tendencies which concentrated 'celebrity' staff and resources in a handful of institutions.

The trend to internationalization in economics teaching and research was a general phenomenon during the last quarter of the century. The diversity, both between and within nations, with which the discipline had begun the century had, by the early post-war period, increasingly given way to homogenization of style and substance. This process accelerated in the 1980s as the personal computer offered every economist access to data and means for its processing without leaving the office. By contrast, most of Bill Phillips's work on inflation and unemployment in the 1950s had been done late at night on the National Physical Laboratory's computer in Teddington. Likewise, Richard Stone had during the 1940s done most of his own statistical work on a hand-cranked machine. The speed with which data could now be processed did away with the enforced lengthy periods during which one pondered the meaning of previous results and devised new strategies. But it also meant that such thinking was at a discount, given the range of data and software. The discipline of economics succumbed to a basic 'law' of markets: the larger the size, the less the diversity.

Nonetheless, public interest in economics survived, and economic careers developed that did not depend upon university status. This new trend originated in the 1980s. Nigel Lawson, Margaret Thatcher's Treasury minister, had a background in economic journalism, symbolizing the rise of a new source of authority independent of any academic institution. Many of the new breed of 'City economist' had no formal academic background in economics at all, but drew upon other technical skills. Independent 'think tanks' began making themselves heard, foremost among them the Institute for Fiscal Studies (IFS), which by the end of the century had grown into the leading non-government authority on domestic fiscal affairs. The rise of the IFS was accompanied by a number of similar organizations addressing the social, political and economic issues that university economics had for the most part left far behind. And finally, a new, non-academic popular literature of economics emerged, seeking to demonstrate the public utility of economic principles to an increasingly receptive readership.

See Also

- ▶ Keynesianism
- ▶ Marshall, Alfred (1842–1924)

Bibliography

Backhouse, R.E. 1996. The changing character of British economics. In *The post-1945 internationalization of economics. Annual supplement to History of political economy*, ed. A.W. Coats, Vol. 28. Durham, NC: Duke University Press.

Backhouse, R.E. 2000. Economics in mid-Atlantic British economics 1945–95. In *The development of economics in Western Europe since 1945*, ed. A.W. Coats. London: Routledge.

Backhouse, R.E. 2006. The Keynesian revolution. In *The Cambridge companion to Keynes*, ed. R.E. Backhouse and B.W. Bateman. Cambridge: Cambridge University Press.

Cairncross, A., and N. Watts. 1989. *The economic section 1939–1961: A study in economic advising*. London: Routledge.

Chester, N. 1986. *Economics, politics and social studies in Oxford, 1900–85*. London: Macmillan.

Clarke, P. 1988. *The Keynesian revolution in the making 1924–1936*. Oxford: Oxford University Press.

Coase, R. 1937. The nature of the firm. *Economica NS* 4: 386–405.

Coats, A.W. 1993a. Some observations on the role of the economist in government. In *The sociology and professionalization of economics. British and American essays*, Vol. 2. London: Routledge.

Coats, A.W. 1993b. Britain: The rise of the specialists. In *The sociology and professionalization of economics. British and American essays*, Vol. 2. London: Routledge.

Coats, A.W. 1993c. The market for economists in Britain, 1945–75: A preliminary survey. In *The sociology and professionalization of economics. British and American essays*, Vol. 2. London: Routledge.

Collard, D. 1981. A.C. Pigou, 1877–1959. In *Pioneers of modern economics in Britain*, ed. D.P. O'Brien and J.R. Wesley. London: Macmillan.
Committee on Higher Education. 1963. *Higher education. Report of the committee appointed by the Prime Minister under the Chairmanship of Lord Robbins 1961–63.* Cmnd. 2154. London: HMSO.
Corry, B. 1997. Bernard Corry. In *Economic careers. Economics and economists in Britain 1930–-1970*, ed. K. Tribe. London: Routledge.
Groenewegen, P. 1995. *A Soaring Eagle: Alfred Marshall 1842–1924*. Cheltenham: Edward Elgar.
Hansen, A. 1953. *A guide to Keynes*. New York: McGraw Hill.
Hicks, J.R. 1937. Mr. Keynes and the classics. *Econometrica* 5: 147–159.
Hicks, J.R. 1939. *Value and capital*. London: Oxford University Press.
Kadish, A. 1982. *The Oxford economists in the late nineteenth century*. Oxford: Oxford University Press.
Kadish, A. 1993. The city, the Fabians and the foundations of the London school of economics. In *The market for political economy. The advent of economics in British university culture, 1850–1905*, ed. A. Kadish and K. Tribe. London: Routledge.
Kaldor, N. 1941. The White Paper on national income and expenditure. *Economic Journal* 51: 181–191.
Klein, L. 1948. *The Keynesian revolution*. New York: Macmillan.
Knight, F. 1921. *Risk, uncertainty and profit*. New York: Houghton Mifflin Company.
Laidler, D. 1999. *Fabricating the Keynesian revolution: studies of the inter-war literature on money, the cycle, and unemployment*. Cambridge: Cambridge University Press.
Meade, J.E. 1936. *Economic analysis and policy*. London: Oxford University Press.
Middleton, R. 1998. *Charlatans or saviours? Economists and the British economy from Marshall to Meade*. Cheltenham: Edward Elgar.
Morgan, M. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.
Peden, G. 1988. *Keynes, the Treasury and British economic policy*. London: Macmillan.
Pigou, A.C. 1912. *Wealth and welfare*. London: Macmillan.
Pigou, A.C. 1920. *The economics of welfare*. London: Macmillan.
Pigou, A.C. 1925. In memoriam. In *Memorials of Alfred Marshall*, ed. A.C. Pigou. London: Macmillan.
Pigou, A.C. 1927. *Industrial fluctuations*. London: Macmillan.
Plant, A., and L. Robbins. 1937. La London School of Economics. In *L'Enseignement Économique en France et à l'étranger,* Special 50th Anniversary Issue of *Revue d'Économie Politique, 66–78.*
Presley, J.R. 1981. D.H. Robertson, 1890–1963. In *Pioneers of modern economics in Britain*, ed. D. O'Brien and J.R. Presley. London: Macmillan.
Reddaway, W.B. 1936. The general theory of employment, interest and money. *Economic Record* 12: 28–36.
Robertson, D. 1948. A study of industrial fluctuation, 1915: New introduction. In *A study of industrial fluctuation.* LSE Reprint No. 8. London.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Rymes, T.K. 1989. *Keynes s lectures, 1932–35*. Houndmills: Macmillan.
Samuelson, P. 1955. *Economics*. 5th ed. New York: McGraw Hill.
Sayers, R.S. 1983. 1941: The first Keynesian budget. In *The managed economy. Essays in British economic policy since 1929*, ed. C. Feinstein. Oxford: Oxford University Press.
Shackle, G. 1967. *The years of high theory: Invention and tradition in economic thought 1926–1939*. London: Cambridge University Press.
Skidelsky, R. 1992. *John Maynard Keynes vol. two. The economist as saviour 1920–1937*. London: Macmillan Press.
Tribe, K. 1997. *Economic careers: Economics and economists in Britain 1930–1970*. London: Routledge.
Tribe, K. 2000. The Cambridge economics tripos 1903–55 and the training of economists. *Manchester School* 68: 222–248.
Tribe, K. 2001. Economic societies in great Britain and Ireland. In *The spread of political economy and the professionalisation of economists*, ed. M. Augello and M. Guidi. London: Routledge.
Young, W. 1987. *Interpreting Mr. Keynes: The IS–LM enigma*. Oxford: Polity Press.
Young, W., and F.S. Lee. 1993. *Oxford Economics and Oxford Economists*. Basingstoke: Macmillan.

British Classical Economics

Mark Blaug

Abstract

Classical economics is not just a period in the history of economic thought immediately prior to the marginal revolution but involves a distinct approach to economic problems. But endless controversy surrounds the definition of that approach. Indeed, the scope of the science of political economy as conceived in Smith's *The Wealth of Nations* was sharply contracted in Ricardo's *Principles of Political Economy.*

Some modern commentators characterize classical economics as surplus theory; others as general equilibrium theory. Economists who are divided in their views will always try to find those views embodied in the writings of the past.

Keywords

Absolute advantage; Austrian economics; Babbage C.; Bailey S.; Balance of payment; Blaug M.; British classical economics; Cairnes J.; Capital accumulation; Capital–Labour ratio; Clark J. B.; Classical economics; Comparative advantage; Corn Laws; Corn model; Cournot A.; Distribution; Dupuis A.-J.-E.; Eagly R.; Exchange value; Factor endowments; Free banking; Free trade; General equilibrium theory; Gossen H.; History of economic thought; Hollander S.; Increasing returns; International trade; Theory of; Invariable measure of value; Jevons W.; Joint production; Jones R.; Keynes J.M.; Knight F.; Labour theory of value; Laissez Faire; Lloyd W.; Longfield M.; Malthus T.; Marginal productivity theory; Marginal revolution; Marshall A.; Marx K.; McCulloch J.; Menger C.; Mill J.; Mill J. S.; Money; Natural price; Neoclassical economics; Neo-Ricardian economics; Normal price; O'Brien D.; Opportunity cost; Organic composition of capital (Marx); Petty W.; Physiocracy; Pigou A.; Political economy; Population growth; Poverty alleviation; Preferences; Pressure groups; Production techniques; Productive vs unproductive labour; Profit and profit theory; Quantity theory of money; Quesnay F.; Rae J.; Rate of profit; Reciprocal demand; Rent; Reserve army of the unemployed; Ricardian revolution; Ricardo D.; Say J.-B.; Say's Law; Schumpeter J.; Senior N.; Smith A.; Sowell T.; Sraffa P.; Standard composite commodity (Sraffa); Static equilibrium analysis; Steady-state equilibria; Subsistence and subsistence wages; Supply and demand; Surplus; Thornton H.; Thünen G.; Tooke T.; Torrens R.; Transformation problem; Use Value; Utility theories of value; Vent-for-surplus doctrine; Wages fund doctrine; Wakefield E.; Walrasian theory of general equilibrium; Wicksteed P.

JEL Classifications

B1

B

The label 'classical economics' is sometimes employed to refer quite simply to an era in the history of economic thought from, say, 1750 to 1870, in which a group of predominantly British economists used Adam Smith's *Wealth of Nations* as a springboard for analysing the production, distribution and exchange of goods and services in a capitalist economy. So broad a definition of classical economics must include such contemporary Continental writers as Cournot, Dupuit, Thünen and Gossen, not to mention such British writers as Bailey, Lloyd and Longfield, who at first glance seem to stand outside the tradition founded by Adam Smith. It is difficult to resist the implication, therefore, that classical economics is more than a period in the history of economic thought: it seems to involve a definite approach to economic problems. The difficulty, however, is how to characterize this approach.

Shrugging aside such tendentious definitions of classical economics as those of Marx and Keynes – for Marx (1867, pp. 174–5n) classical political economy begins with Petty in the 17th century and ends with Ricardo, and for Keynes (1936, p. 3n) the classical school begins with Ricardo and ends with Pigou – the first question is whether it was Adam Smith or David Ricardo who established the 'essence' or 'core' or classical economics. Of course, Adam Smith laid down the main issues that economists debated for a century after him, but there is also little doubt that the Smithian tradition was in some sense transformed with the appearance of Ricardo's *Principles of Political Economy and Taxation* in 1817. Some writers have nevertheless insisted that Smith and not Ricardo was the lasting influence on the character of classical economics, contending that the leading features of Ricardo's theoretical system were soon rejected even by his avowed followers in the decade after his death in 1823. Others, however, have insisted that, despite all the criticisms of Ricardo that no doubt appeared in the late 1820s and early 1830s, later writers like John Stuart Mill and John Elliott Cairnes continued to

operate right up to the 1870s with the central Ricardian theorem that the rate of profit and hence the accumulation of capital depends critically on the marginal cost of production in agriculture; in that sense, they remained trapped in the Ricardian system. But even this assertion presupposes the notion that the Ricardian system is essentially characterized as a theory about the determination of the rate of profit, a proposition which is by no means accepted by all historians of economic thought.

It is only after clearing up this problem of the relative significance of Smith's and Ricardo's ideas in shaping the central current of classical economics that we can take up the question of where to place the utility theories of value put forward by such writers as Lloyd, Longfield, Senior, Dupuit and Gossen, the abstinence theories of interest of Bailey, Senior, Rae and John Stuart Mill, the use of both supply and demand forces in the determination of international prices by Mill, the theory of general gluts and the denial of Say's Law of Markets by Malthus, and the exploitation theory of profits by Marx – in short, all the elements of economic theorizing in the period 1770–1870 that so clearly do not belong to the corpus of doctrines bequeathed by Adam Smith and David Ricardo. Likewise, it is only then that we can start talking about the end of classical economics in the 1870s and the nature of the 'marginal revolution' that may or may not have marked a decisive break in the continuity of orthodox economics.

The endless debate on what was classical economics is neatly illustrated by the simultaneous appearance of three books on classical economics: *Classical Economic Reconsidered* by Thomas Sowell (1974), *The Structure of Classical Economic Theory* by Robert Eagly (1974) and *The Classical Economists* by Denis O'Brien (1975). Of the three, Eagly takes the widest view of the length of time over which something called 'classical economic theory' ruled the roost, beginning with the physiocrats in the 1750s and ending with the Walrasian theory of general equilibrium in the 1870s. His view is not only that the whole of classical economics can be defined in terms of a single conceptual framework but that this framework revolves essentially around a particular concept of capital as a stock of intermediate goods invested in staggered production periods, the question of the pricing of final goods always relegated to the next period after output has already been determined by the size of the labour force and the technology of the previous period; in short, the key to classical economics is to be found in the so-called 'wages fund doctrine'. Whether this thesis is convincing or not, Eagly's book represents an extreme example of the tendency to define classical economics as one coherent body of ideas organised around a central unifying principle. The secondary literature is, of course, replete with other attempts to pin down once and for all the classical theory of economic growth (e.g. Lowe 1954; Samuelson 1978), but few allege, as Eagly does, that their modelling of classical economics captures all the essentials of the writings of Quesnay, Smith, Ricardo, Mill and Marx, as well as McCulloch, Torrens, Bailey, Jones, Senior, Longfield, Babbage, Tooke, Wakefield, etc.

Sowell, on the other hand, adopts the traditional definition of classical economics as in effect the School of Adam Smith, and he therefore excludes Marx and, more surprisingly, Malthus, Torrens and Senior at least in some respects from the mainstream of the tradition stemming from *The Wealth of Nations.* That tradition consisted, according to Sowell, of a common set of philosophical presuppositions, common methods of analysis and common conclusions regarding matters of substantive economic analysis: it comprised such major propositions as the labour theory of value, the Malthusian theory of population, Say's Law and the quantity theory of money and was predominantly oriented towards the issue of economic growth (although not in the modern sense of the term as a theory of the steady-state equilibrium growth path of an economy). However, Sowell admits that this picture has to be qualified after 1817 by such phrases as 'classical economics in its Ricardian form' because Ricardo worked a major change in Smith's eclectic mode of economic reasoning by adopting static equilibrium analysis as the only valid method of conducting an economic argument. At any rate,

Sowell's treatment of classical economics leaves little doubt of the extensive and varied character of economics in the classical period, posing problems for anyone who seeks to define classical economics in one or two sentences.

Both Eagly's and Sowell's books are dwarfed by O'Brien's wide-ranging and comprehensive review of classical economics, which alone among the three begins with an incisive discussion of the extent to which the classical writers formed a 'scientific community'. (O'Brien's book also contains excellent annotated bibliographical notes on classical economics; indeed, O'Brien, Blaug (1985) and Spiegel (1983) between them review the whole of the secondary literature.) O'Brien follows Schumpeter in arguing that the Ricardian system represented an analytical detour from the main line of advance running from Adam Smith to John Stuart Mill; it was not a fatal detour, however, because the full Ricardian apparatus attracted hardly any followers and in any case was more or less abandoned by the 1830s. As we noted earlier, this Schumpeter–O'Brien thesis has been questioned by some (e.g. Blaug 1958; Hollander 1977). The point is, however, that O'Brien's book perfectly illustrates our contention that any stand taken on the nature of classical economics as a whole depends critically on the attitude adopted towards the Ricardian metamorphosis of Smithian economics.

The Sraffa Interpretation of Ricardo

Still more recently a new note has been struck in the old argument about the essential meaning of classical economics. Inspired by the publication of Sraffa's *Production of Commodities by Means of Commodities* (1960), a number of commentators have argued that classical economics is in effect a Sraffa-system, that is, an analysis of the manner in which a capitalist economy invests its surplus of net output over consumption, which is to say an output in excess of that required to reproduce that level of output, subject to the condition that goods and services are so priced as to maintain a uniform rate of wages and a uniform rate of profit on capital in all lines of investment. This approach, they contend, was buried in the 1870s when the central object of economic analysis became that of investigating the optimum allocation of resources whose quantities are given at the outset of the analysis; in reviving classical surplus analysis, Sraffa not only provides a promising new way of studying economic problems but also illuminates precisely what it was that united Smith, Ricardo and Marx, thus licensing the use of a single label such as 'classical economics' to cover them all (see Meek 1973, 1977, the originator of the argument; and Dobb 1973; Roncaglia 1978; Walsh and Gram 1980; Bradley and Howard 1982; Eatwell 1982; Garegnani 1984; Howard and King 1985).

As is well known, a Sraffa-system consists of a set of linear production equations, one for each commodity in the economy, and is intended to demonstrate that these equations are sufficient to determine all relative prices in long-run equilibrium irrespective of the pattern of demand, provided that (1) the output of each commodity is given; (2) rate of profit on capital is uniform throughout the economy and (3) the real wage or (alternatively the rate of profit on capital) is somehow determined exogeneously. On the face of it, such a theory does indeed appear to be very much like 'classical economics'. For example, after distinguishing between 'natural' and 'market' prices of commodities – or, as we would nowadays say, the long-run and short-run prices of commodities – Adam Smith focused much of his analysis on the determination of 'natural' prices, a tendency which became even stronger in the writings of Ricardo. Moreover, Smith and certainly Ricardo, not to mention Marx, always wrote as if demand played no role whatever in the determination of 'natural' price. We have all known ever since the work of Marshall that this neglect of demand can be justified if one assumes that commodities are produced under conditions of constant unit costs or constant returns to scale, the long-run supply curves of all industries being perfectly horizontal over the relevant range of output. Sraffa's production equations imply fixed coefficients of production and, again, we have known ever since the work of Leontief that fixed coefficients of production are sufficient (but not

necessary) to produce constant costs. In short, Sraffa's demonstration that prices in his model are determined independently of demand is eminently 'classical'.

Likewise, there is no doubt that the concept of a uniform rate of return on capital, or rather defining 'natural' prices to be those generated by a stationary equilibrium in which the rate of profit has become equalized by interindustry mobility of capital, is typical of all economic writing in the century between 1770 and 1870. Finally, the real wage rate in classical economics is determined by so-called 'subsistence' requirements and these were defined by Ricardo, Mill and Marx in historical rather than physiological terms; in other words, it was assumed that the current 'natural' price of labour reflected the past history of the 'market' price of labour. The 'natural' price of labour was in effect determined by workers' attitudes to the size of their families but since the classical economists did little to analyse these attitudes, it is not too much to say that the so-called 'subsistence *theory* of wages' actually amounts to taking 'subsistence' as a datum (Schumpeter 1954, p. 665). Once again, it can be argued that the Sraffian assumption of an exogeneous real wage is 'classical' in spirit.

There is no doubt that Sraffa's system captures many of the elements of 'classical economics'. It provides a further bonus, however, in illuminating classical economics. Generations of critics have tried to make sense of Ricardo's lifelong quest for an 'invariable measure of value' and have given it up as a hopeless task. Ricardo was troubled by the fact that any change in money wages will alter the structure of relative prices owing to the fact that capital and labour are combined in different proportions in different industries. Thus, a rise in wages or a fall in the rate of profit raises the prices of labour-intensive goods relative to the price of capital-intensive goods. This violates the labour theory of value according to which relative prices are determined by the physical quantities of labour expended on production independently of the rate at which labour is rewarded. To remedy this difficulty, Ricardo struck upon the notion of expressing all prices in terms of a commodity produced by a ratio of capital to labour that is a weighted average of the entire spectrum of capital–labour ratios in the economy; such a commodity, he believed, constitutes an 'invariable measure of value' in the sense of providing a standard of measurement that is invariant to changes in the ratio of wages to profits. In the same way, Sraffa measures all prices in terms of a 'standard composite commodity' that consists only of outputs combined in the same proportions as the non-labour inputs that enter into all the successive layers of its manufacture. Moreover, in one of the many elegant demonstrations in his book, Sraffa succeeds in showing that such a 'standard commodity' is in fact embedded in any actual economic system and that the proportion of net output going to wages in that reduced-scale system determines the rate of profit in the economy as a whole.

The explanation of this result depends on Sraffa's distinction between 'basic' commodities which enter directly or indirectly into the production of every commodity in the economy, including themselves, and 'non-basic' commodities which enter only into final consumption. If we treat labour itself as a produced 'means of production' then wage goods constitute examples of 'basic' commodities, that is, they are technically required to cause households to produce the flow of labour services. Ricardo clearly believed that wheaten bread was 'basic' in this sense but Sraffa parts company with Ricardo in rejecting any and all versions of the subsistence theory of wages; workers in Sraffa are primary, non-reproducible inputs. Nevertheless, there are plenty of other basics besides wage goods in an actual economy and the upshot of Sraffa's distinction between basics and non-basics is that the 'standard composite commodity' consists only of basics and indeed of all the basics in the economy; this collection of basics enters into the production of the invariant yardstick in a 'standard ratio', that is, in the same proportion as they enter into their own production. It turns out that relative prices and either the rate of profit or the rate of wages (depending on which one is given exogeneously) depend only on the technical condition of producing the 'standard commodity' and are in no way affected by what happens to nonbasic

commodities. In a way this is obvious: a change in the cost of producing a nonbasic no doubts alters its own price but, by the definition of a nonbasic commodity, the effect stops there since the product in question never becomes an input into any other technical process. It is also obvious, at least intuitively, that an exogenous change in wages unconnected with a change in productive techniques alters the rate of profit but has no effect on relative prices measured in terms of the standard commodity for the simple reason that the change alters the measuring rod in the same way as it alters the pattern of prices being measured. The 'standard commodity' therefore provides an 'invariable measure of value', and Ricardo's old problem is at long last solved.

In developing his own ideas, Sraffa also advanced an entirely new interpretation of how Ricardo came to connect his theory of the determination of the rate of profit with the question of finding an invariable yardstick for measuring relative prices. In his early pamphlet *Essays on the Influence of a Low Price of Corn on the Profits of Stock* (1815), Ricardo wanted to show that the extension of cultivation to inferior soils depresses the rate of profit on capital throughout the economy by raising the marginal cost of producing 'corn', that is, wheat, the principal wage good consumed by workers. This is easy to demonstrate in a one-sector economy where the only output is wheat. However, from the beginning Ricardo operated with a two-sector economy in which an agricultural industry produces 'corn' and a manufacturing industry produces 'cloth'. Of course, if wage goods consist entirely of corn and if cloth is always purchased out of profits and rents, it is still easy to show that the rate of profit on capital depends decisively on the action of diminishing returns in agriculture. In agriculture, wheat is the only output and it is also the input both in the form of wages 'advanced' to workers to tide them over the annual production cycle and seeds to plough back into the next agricultural cycle; hence, the 'money' rate of profit in agriculture cannot possibly diverge from the 'wheat' rate of profit because any change in the price of wheat affects inputs and output in the same degree. Manufacturing, however, only uses wheat as one of its inputs (namely, in the form of wage goods), and since the rate of profit earned on capital must be equal in between the two industries in equilibrium, the price of wheat determines a definite price for cloth. If, for example, the rate of profit in agriculture falls due to the operation of diminishing returns, the price of cloth in terms of wheat must likewise fall to prevent cloth from being more profitable to produce than wheat. To reiterate: measuring all prices in terms of wheat, the 'money' rate of profit in industry is governed by the 'wheat' rate of profit in agriculture, which, in turn, depends entirely on the technology of producing wheat, the unique wage good; in one of Ricardo's famous catch phrases: 'it is the profits of the farmer which regulate the profits of all other trades'.

This ingenious argument, which appears to explain the determination of the rate of profit in purely physical terms without the use of a theory of value, is known in the literature as the 'corn model'. In the preface to his edition of *The Works of David Ricardo* (1951), Sraffa argued that the corn model is implicit in Ricardo's 1815 *Essay*. To be sure, Ricardo never wrote it down in so many words because even in the *Essay* he could not swallow the assumption that wages are entirely spend on wheat, that all agricultural products are wage goods and that all manufactured products are luxuries which are never consumed by workers. Nevertheless, he did use wheat in the *Essay* as a measure for aggregating the heterogeneous inputs of agriculture on the assumption that all prices rise and fall with wheat prices, and he also employed arithmetical examples in which all inputs and outputs of both agriculture and manufacturing are expressed in terms of wheat. In the *Principles* he analysed an economy with many sectors in which a change in the terms of trade between wheat and cloth will alter real wages and hence the rate of profit on capital. Nevertheless, his preoccupation in this mature work with the 'invariable measure of value' may be read as an attempt to secure the same results obtained earlier with the aid of the corn model, that is, to tie the determination of the rate of profit directly to the production function of agriculture. Of course, if Ricardo could have ignored the

varying proportions of labour and capital in different industries, he could have reached all his conclusions without the aid of an invariable yardstick of value. He had placed so much emphasis, however, on what Marx was to call the unequal 'organic composition of capital' that this route was closed to him. Hence, the quest for an 'invariable measure' with which to recapture the simple truth of the corn model. Here then is a rational reconstruction of Ricardo's arguments that accounts neatly for both the form and the drift of his reasoning.

A General Equilibrium Interpretation of Ricardo

Sraffa's interpretation of Ricardo has won wide assent even among those who otherwise remain sceptical about Sraffa's system in its own right. However, Samuel Hollander's recent reexamination of the whole of Ricardo's writings has taken sharp exception to Sraffa's reading (Hollander 1979, pp. 123–90, 684–9). Ricardo, according to Hollander, never entertained the corn model even implicitly, never assumed that corn alone enters the wage basket, never argued that the rate of profit in agriculture determines the general profit rate and, above all, never assumed that real wages remain constant either because they are determined by the subsistence requirements of workers or because they are determined exogenously. What Hollander really objects to is the notion that 'distribution', that is, the rate of wages and the rate of profit, are determined in Ricardo as in Sraffa's own model independently of and indeed prior to the value of commodities, so that the former causally determines the latter. This is to be contrasted with the approach of Walrasian general equilibrium theory in which the pricing of factor services is determined simultaneously with the pricing of final consumption goods. It is simply not true, argues Hollander, that the history of economic thought can be neatly divided into two great branches, a general equilibrium branch leading down from Walras and Marshall to Samuelson, Arrow and Debreu today, in which all relevant economic variables are mutually and simultaneously determined, and a completely different branch leading down from Ricardo and Marx to Sraffa in which distribution takes priority over pricing because economic variables are causally determined in a sequential chain starting from a predetermined real wage (Pasinetti 1974, pp. 42–4, even enlists Keynes into the ranks of the Ricardo–Marx–Sraffa school). Ricardo, Hollander insists, was essentially a general equilibrium theorist – and so were Adam Smith, John Stuart Mill and even Karl Marx (Hollander 1973, 1981, 1982).

Before passing judgement on this dispute, it is worth nothing that what has been called the 'neo-Ricardian' or 'Cambridge' interpretation of the history of economic thought claims superior merit for Ricardo because Ricardo divorced the question of distribution from the question of pricing. But this is precisely the grounds on which many pre-war historians of economic thought attacked Ricardo! Thus, Frank Knight in a famous essay on 'The Ricardian Theory of Production and Distribution' (1956) poured scorn on classical writers like Ricardo because they utterly failed to approach the problem of distribution as a problem of valuation and this despite the fact that the effective demand for any factor of production depends on the distribution of income, which in turn depends at least to some extent on the pricing of factor services; in short, 'distribution theory has little meaning apart from a theory of general equilibrium' (Knight 1956, pp. 41, 63). Similarly, Schumpeter (1954, pp. 473, 568–9, 1171) spoke scathingly of the 'Ricardian Vice' whereby an already oversimplified economic model is further reduced by freezing one endogeneous variable after another by special *ad hoc* assumptions. First, rent in Ricardo is determined as an intramarginal return to land treated as a factor in fixed supply; the location of the margin depends of course on the demand for agricultural produce, but this is in turn explained by the size of the population via the assumption of a perfectly inelastic demand for corn. Second, having 'gotten rid of rent' on the margins of cultivation, Ricardo then employed a subsistence theory of wages to determine the share of total-output-minus-rent that accrues to labour. Third, total profits in

Ricardo are treated as a pure residual after the deduction of wages and rents, the rate of profit being determined as the quotient of total profits and the inherited stock of capital. In other words, the problem of distribution is explained by three totally different types of theories, which in turn are quite different from the principles employed to explain the pricing of goods and services, namely, the labour theory of value. How amazed Knight and Schumpeter would have been to see their critique stood on its head, so that what they regarded as vices are now viewed in certain quarters as virtues.

Ricardo Versus Smith

Having expounded various interpretations of classical economics, it is time to attempt some sort of general assessment. To collect our thoughts, consider the number of problematic issues we have outlined above. Is the economics of Adam Smith something different from the economics of David Ricardo? Obviously there is no total break in the continuity of thinking, but nevertheless, is there a sufficient break to warrant the use of such dramatic language as the 'Ricardian Revolution'? Was this 'Ricardian Revolution' the implicit resort to something like the 'corn model' to produce a clear-cut explanation of the determination of the rate of profit, or was it simply a change in the style of economic reasoning? Was Ricardo soon repudiated, so that the Smithian tradition survived right down to John Stuart Mill and beyond, or are the later phases of classical economics dominated by the ideas of Ricardo rather than those of Adam Smith? Is there sufficient coherence around a definite core of ideas to permit us to talk at all of 'classical economics'? Is this core the notion of the origin and disposition of the 'economic' surplus and the proposition that distribution is independent of valuation? And, finally, is all of classical economics a primitive but prescient version of general equilibrium analysis?

We can deal quickly with the first question, the so-called 'Ricardian Revolution'. With the exception of Hollander (1979, ch. 1), all modern commentators on classical economics agree that Ricardo altered the scope, method and focus of economics. Even if we take only *The Wealth of Nations* among Smith's books and essays, the scope of economics for Adam Smith is enormous and perhaps wider than that for any economist before or after him. The first two books of *The Wealth of Nations* consists largely of what later came to be regarded as the very hallmark of orthodox economics: the theory of value and the theory of production and distribution, employing in the main the method of comparative statics. But even the 'Digression' on the value of silver in chapter 11 of Book I takes up an unorthodox topic, namely, changes in the structure of prices over centuries with the aid of a method of analysis that might be called 'inductive' or 'historical'. Moreover, here as elsewhere in *The Wealth of Nations* there is a remarkable emphasis on the notion of 'increasing returns' so widely defined as to include the effects of both increases in the scale of production and changes in the method of production or technical progress. Despite the flowering of a considerable literature in recent years purporting to model Smith's 'theory of economic growth', few have succeeded in capturing this vital element in Smith's thinking, which Kaldor (1972) has consistently emphasized (but see Eltis 1984, ch. 3). Moreover, this notion of increasing returns soon dropped out of classical economics, coming back only ninety years later with the writings of Karl Marx.

Similarly, there is the famous distinction in Book III of *The Wealth of Nations* between productive and unproductive labour which Ricardo and Mill accepted, which McCulloch and Senior denied, which Marx reinterpreted in a different way, but which nevertheless was never followed up and developed in any fruitful way. A simple explanation for this failure to elaborate Smith's distinction was that Smith made a mess of it, defining productive labour alternatively as labour which produces something tangible, produces a profit for its employer, and generates productive capacity that then creates a demand for additional employment. But another explanation is that the distinction between the employment of 'manufacturers' and 'menial servants', between wealth-creating and wealth-consuming activities, is only

relevant in the context of long-run economic development, being partly a 'positive' account of different patterns of economic change in different nations and partly a 'normative' proposal for legislators seeking to maximize the rate of net investment in an economy. Although Mill was profoundly concerned with questions of economic development (see O'Brien 1975, ch. 8), Ricardo had no real interest in the forces that govern the historical patterns of economic change, and for that reason alone the Smithian distinction between productive and unproductive labour, and the associated discussion of an optimum investment pattern between industries in chapter 5 of Book II of *The Wealth of Nations*, was effectively laid to rest all through the heyday of classical economics.

Smith's interest in 'the different progress of opulence in different ages of nations' totally dominates Book III of *The Wealth of Nations* and is at work even in Book IV on mercantilist theory and policy and Book V on public finance. In this latter half of *The Wealth of Nations* there is little appeal to the comparisons of steady-state equilibria, which was to figure so heavily in practically everything that Ricardo wrote. But there are two other elements in these pages that are totally missing in Ricardo and even in Mill, namely, a concern with the incentive effects of different institutional devices for rewarding self-employed professionals and individuals employed in the public sector (Rosenberg 1960) and a keen sense of the role of pressure groups in the formulation of economic policies (Peacock 1975; West 1976; Winch 1983). Thus, the modern theory of property rights as well as the economic theory of politics may properly claim Smith as a forerunner. At any rate, neither of these two aspects of *The Wealth of Nations* has any echoes in the writings of those that came immediately after Smith.

Consider next the theory of international trade. There is a static equilibrium theory of the gains of foreign trade in Smith based on the principle of absolute rather than comparative advantage, and here no doubt, Ricardo saw further than Smith. But there is also a dynamic theory of the gains of trade in Smith, the so-called 'vent-for-surplus' doctrine, according to which foreign trade widens the extent of the market and generates new wants; this view of foreign trade disappears in Ricardo and only comes back to classical economics with Mill (Bloomfield 1975, 1978, 1981).

Smith's theory of money is also profoundly different from that of Ricardo, typically invoking the quantity theory of money in its dynamic 18th-century version in which the emphasis falls on the disequilibrium 'transition period' between an increase in the quantity of money and the rise in prices and not on the final equilibrium adjustment between money and prices (Laidler 1981). In addition, Smith was an advocate of private, unregulated banking (qualified only by the prohibition of the issue of banknotes for small sums), reflecting the operation of Scottish banking, which was unregulated for over a century between 1716 and 1844. It was Henry Thornton who first rejected the Smithian tradition in his *Paper Credit of Great Britain* (1802), explicitly denying that the note issue in a free banking system would be self-regulating as Smith had argued. By the time of Ricardo it was orthodox to argue that the issue of banknotes was an obvious exception to the doctrine of laissez faire (White 1984, ch. 3). Here too, the gulf between Smith and Ricardo is almost total.

There is no need to underline Ricardo's differences with Adam Smith over the labour theory of value, since Ricardo set out explicitly to criticize Smith's failure to apply the labour theory of value to a modern economy rather than a purely conjectural 'early and rude state of society'. But what is not so obvious is the fact that even in respect of labour as a measure of the 'real price' of commodities – Smith's tortured language in Book I, chapter 5, for the problem of specifying an index number of economic welfare – Smith's view of labour is profoundly subjective, whereas Ricardo in his comparable chapter 20 of the *Principles of Political Economy and Taxation* on 'value and riches' consistently treats labour as an objective, physical expenditure of energy. In the masterly tenth chapter of Book I of *The Wealth of Nations* on 'relative wages', Smith demonstrated that competition in labour markets equalize the net advantages of different occupations, that is, the monetary returns to units of disutility of labour. In other words, to the extent that labour

is a 'measure of value' in Smith, it is labour conceived as 'toil and trouble' and reflects the preferences of workers as much as those of their employers. Although Ricardo, and for that matter Marx, never disputed this analysis of Smith, they ignored its implications and blithely treated labour as fundamentally homogeneous in quality, its role in the production of commodities being conceived as a brute reflection of purely technological data; in short, they took as given something like Sraffa's production equations. It is this and not the famous debate over whether the value of commodities in Smith is determined by the labour 'commanded' by goods or the labour 'embodied' in their production that represents the real watershed in the history of the labour theory of value (Robertson and Taylor 1957; Gordon 1959; Blaug 1985, pp. 49–53).

But the most profound departure in Ricardo from the Smithian tradition is the notion that rent is in a class by itself as a source of income: it is 'unearned income', being an intramarginal return to purely natural differences in the quality of land which have nothing whatever to do with the activity of landlords. Despite Smith's references to landlords who 'love to reap where they have never sowed' and the 'conspiracy' of merchants, the Smithian world is one in which all economic interests are essentially harmonious or, at any rate, capable of being made harmonious by wise legislators. The Ricardian world, however, is one which conflicting class interests are unavoidable. It is this unique element in the Ricardian system, which gave classical economics its sharp political edge, an edge that clearly worries so many of the minor classical economists, such as Jones, Senior and Longfield.

Finally, the central and indeed sole focus of the Ricardian system is the question: what determines the rate of profit on capital, or rather, what governs its changes over time? This is a question which never really troubled Adam Smith. He made it clear that profit is equalized among industries in the long run, but he had no explanation of how the level of the rate of profit is determined. To be sure, Smith believed that the rate of profit was eventually doomed to fall because of the exhaustion of profitable investment outlets. But he never emphasized this proposition and on balance he took an extremely optimistic view of the feature prospectus for economic growth. Ricardo too was essentially an optimist about the long-run growth potential of the British economy but only if the Corn Laws were repealed; he was thus motivated to argue the strongest possible connection between the rate of profit on capital and the real cost of producing wheat exclusively with domestic resources. In consequence, Ricardo viewed absolutely every aspect of economic activity, including monetary forces, currency arrangements, taxation, the financing of the public debt, and of course foreign trade, through the lenses of his theory of profits. Many readers of Ricardo have been deceived by the preface to his *Principles* – 'To determine the laws which regulate this distribution (of rent, profit, and wages), is the principal problem in Political Economy' – into believing that the Ricardian system is largely devoted to an analysis of the determination of the relative shares of land, capital and labour. But while Ricardo certainly had much to say about the issue of relative shares, and indeed was responsible for introducing this theme into economics, his analysis is in fact concentrated on rents per acre, the rate of the profit per unit of capital and the rate of wages per man. It is, in a word, a book about the pricing of factor services and that is (surely?) much less than the subject-matter of *The Wealth of Nations.*

There is little doubt, therefore, that the scope of the science of political economy as conceived in *The Wealth of Nations* was sharply contracted in Ricardo's *Principles of Political Economy.* But, in addition, Adam Smith wrote much besides *The Wealth of Nations.* Quite apart from *The Theory of Moral Sentiments* and the remarkable essay on the *History of Astronomy*, the publication of the new University of Glasgow edition of the complete *Works and Correspondence of Adam Smith* (1976–83) strongly suggests that he intended to round off his contributions by a major work on the theory of jurisprudence which he never lived to write; nevertheless, even in *The Wealth of Nations* he never lost sight of the fact that political economy may be considered as 'a branch of the science of a statesman or legislator', the latter being

therefore something more comprehensive than the former. A number of recent commentators (Cropsey 1957; Lindgren 1973; Winch 1978; Skinner 1979) have indeed insisted that all of Adam Smith's writings are held together by a unified vision of an all-embracing social science, which he unfortunately never succeeded in realizing to the full. Whether this thesis is persuasive or not, it certainly strengthens the contention that the economics of Adam Smith is conceived on grander lines than the economics of David Ricardo.

The Corn Model again

So there was what might be described in highly coloured language as a 'Ricardian Revolution': what began as a criticism of some of 'Professor Smith's opinions' ended up as a wholesale revision of the legacy of Adam Smith.

What was the cornerstone of this 'Revolution'? Was it the 'corn model'? It certainly was a denial of the Smithian cost-of-production theory according to which a rise in money wages would raise all prices, thus leaving the rate of profits unaffected. But that is not to say that Ricardo's fundamental theorem that 'profits vary inversely as wages' was based on an implicitly held corn model. It is true that the corn-model interpretation neatly rationalizes Ricardo's arguments in the early *Essay on Profits* in which the economy is conceived as consisting of two sectors but the rate of profit is determined exactly as it would be in a one-sector economy. In other words, Ricardo should have held the corn model for without it the *Essay* is simple logically inconsistent. Nevertheless, the corn-model version simply attributes far more rigour and consistency to Ricardo's analysis than is warranted (Peach 1984). What Ricardo later put in place of the missing corn model was the 'invariable measure of value' which was designed to surmount two of his unresolved difficulties at one and the same time: (1) that workers consume both manufactured and agricultural goods, so that one can never be sure that the rising cost of producing wheat is directly transmitted to the rate of profit; and (2) that capital and labour combine in different proportions in different industries, so that a change in real wages for any reason whatsoever alters the structure of prices and, thus, affects the rate of profit even if nothing has happened to the technology of agriculture.

We noted earlier that Sraffa's *Production of Commodities by Means of Commodities* may be said to have vindicated Ricardo's belief in the existence of an 'invariable measure of value', capable of separating and measuring the effects of changes in technology from those due to changes in the rate of wage and profits. But doubts remain about the validity of this claim. In Ricardo, the divining rod of the invariable measure is supposed to be invariant (as Ricardo kept saying) not just to changes in wages in profits but also to changes in its own methods of production. Sraffa's 'standard commodity' fills the bill on the first score but fails on the second score: it is not invariant to changes in its own techniques of production and therefore falls short of solving Ricardo's problem of linking the determination of the rate of profit directly and unambiguously to the action of diminishing returns in agriculture. The truth is that there is no such thing as an 'invariable' yardstick that will satisfy all the requirements that Ricardo placed upon it (Ong 1983). All of which is to say that, despite the fact that Ricardo was the first truly rigorous analytical economist, it is impossible to exonerate him from all analytical errors: he was at times inclined to square a circle using only a ruler and a compass!

Classical Economics as Surplus Theory

We turn next to the thesis that classical economics is the economics of the creation and disposition of surplus output over consumption – a theory of the reproducibility of economic systems in the making – in sharp contrast to the later neoclassical theme of the allocation of *given* resources between competing ends, subject to the constraints of technology and existing property rights. Now, there can be little doubt that this is precisely the nature of the economics of physiocracy (Eltis

1984, ch. 2), and it is little wonder that those who argue the surplus interpretation include the physiocrats in classical economics (Walsh and Gram 1980, ch. 2). There is also little doubt that it captures much of the drift of *The Wealth of Nations* and turns up again in Mill's *Principles* and in Marx's *Capital.* On the other hand, it does not begin to do justice to dominant features of the Ricardian system and leaves out almost as much as it manages to include in the writings of the classical economists.

What does it tell us, for example, about the jewel in the crown of classical economics: Ricardo's law of comparative advantage as the foundation of the belief in free trade, which served throughout the whole of the 19th century as the litmus-paper test of an economic liberal? Ricardo treated foreign trade as a matter of moving along a static world production-transformation curve, constructed on the basis of *given* resources and the *given* techniques of production of the trading countries; the gains of foreign trade in his celebrated cloth–wine example show up in a global increase in physical output from given labour resources in Portugal and England. There is no hint here of 'surplus theory' and perhaps that is why the surplus interpretation of classical economics studiously avoids discussion of the theory of international trade.

It might be argued, however, that the subject of foreign trade lies outside the mainstream of classical economics because it violates the assumption of a uniform rate of profit on capital – if capital were mobile between countries, international trade would be based like intranational trade on absolute cost advantages. As a matter of fact, Thweatt (1976) has argued that Ricardo's view of foreign trade never went beyond the conception of absolute advantage and this despite the three-paragraph illustration of comparative advantage in his *Principles*, which may well have been written by James Mill rather than Ricardo. After all, free trade for Ricardo meant a policy appropriate to an advanced manufacturing nation in its relation with agrarian nations supplying it with food; the point of the chapter on foreign trade in the *Principles* is not to explain the gains of trade but to demonstrate that foreign trade only affects the rate of profit insofar as it leads to the importation of cheaper wage goods.

Be that as it may, less than a decade after the death of Ricardo, the young Mill (1844, but written in 1829) completed Ricardo's argument by showing that the division of the overall gains from foreign trade in the two countries depends on 'reciprocal demand', thus putting another nail in the coffin of the labour theory of value: even when goods are produced by labour alone within countries, the barter terms of trade between countries depend on both demand and supply. Cairnes subsequently extended the reciprocal demand approach even to domestic trade at least in respect of exchange between 'non-competing groups'. None of this has anything to do with the creation, accumulation and allocation of an economic surplus, and so the surplus interpretation must leave to one side the classical theory of international prices, the classical theory of balance of payment adjustments and with it the classical theory of monetary management.

But the shortcomings of the surplus interpretation extend even to classical theorizing about the operations of a closed economy. It can throw no light on the care with which Adam Smith spelt out the effects of a public mourning on the price of black cloth in Book I, chapter 7, of *The Wealth of Nations*, so as to demonstrate that 'market' prices cannot permanently diverge from 'natural' prices because they imply profit opportunities for producers that will sooner or later be exploited; all this is to say that the surplus interpretation has little time for those short-run adjustments that formed the staple of much of the practical wisdom of classical economists grappling with day-to-day economic problems. Similarly, the surplus interpretation must pass over the doctrine of opportunity costs that was part and parcel of the legacy of Adam Smith, namely, that effective costs to producers are not expenditures incurred in the past but present opportunities foregone. As Buchanan (1929) showed many years ago, Ricardo's characteristic doctrine of 'getting rid of rent' by concentrating attention on the rentless margin of production implies that land has no uses alternative to the growing of wheat; while this may at a pinch be justified at a macroeconomic level,

Smith's theory of rent, which recognizes the fact that land employed in cultivation must compete with land for grazing or urban use, is thus more truly in the tradition of analysing allocation with given resources than is Ricardo's. This Smithian emphasis on the competing uses for land, so that ground rent does enter into the price of agricultural goods, was never lost sight of by classical writers between Ricardo and Mill and comes back into its own in Mill's *Principles*, notably in Book II, chapter 16, on rent theory.

The surplus interpretation is thus a limited view of classical economics, but it is not a misrepresentation. In one sense it is only fancy language for the old view that classical economics is essentially the economics of development, which starts from a fundamental contrast between augmentable labour and non-augmentable land given in quantity and asks how, under these circumstances, growth in the sense of per capita income can be maximized (Myint 1948). Indeed, the notion that growth of population and the accumulation of capital are the great themes of classical economics in contrast to the question of efficient allocation of given supplies of the factors of production in neoclassical economics after 1870 is endorsed in many, if not in all, textbooks on the history of economic thought (e.g. Blaug 1985, pp. 295–6). So why all the fuss? Why all this insistence on the surplus interpretation in recent years?

A close reading of those who have advocated a reading of classical economics in terms of surplus analysis suggests two rather different motivations for the 'new' interpretation: one is to provide Marx with a respectable pedigree, or at least to display Marx as the true heir of bourgeois economics in its days of glory, solving the riddles that that baffled Quesnay, Smith and Ricardo; the other is to reveal Sraffa as the true heir of the classical tradition, demonstrating that there is an old and venerable tradition of explaining the determination of prices without resorting to the preferences and satisfactions of consumers and without relying on a market mechanism to price both capital and labour. Each of these two strands of the surplus interpretation produces its own special distortions of classical economics.

It is certainly true that Marx was in many ways a direct descendant of Smith and Ricardo, and particularly of Ricardo. He took over from Smith the distinction between use value and exchange value (as well as the denial that the former had anything to do with the determination of the latter), the distinction between market and natural prices, together with the notion that the business of the economist is to explain natural prices as terminal states of long-run equilibrium outcomes, the distinction between productive and unproductive labour, the conception of historically increasing returns as a major force in the process of development, the tripartite division of national revenue into wages, profits and rents as the incomes of three distinct social classes – and much else. But he learned even more from Ricardo, and particularly Ricardo's discovery that all the problems of the labour theory of value are reducible to the undeniable fact that capital and labour combine in different proportions in different industries, difficulties which may be resolved however by measuring all prices in terms of the price of a commodity produced by the 'average' industry. This was the key to Marx's 'transformation problem', which demonstrated that 'prices of production' must systematically diverge from labour 'values' if the rate of profit is to be uniform between industries, an insight which, Marx thought, had always eluded Ricardo. Marx hardly noticed that in correcting Ricardo's answer, he also corrected his question. Ricardo's problem had been: what determines the rate of profit? Marx's problem, however, was: what determines the rate of profit if profit is in the nature of unpaid labour, a mark-up on the outlays of wages disguised as a mark-up on all cost-outlays? But the nature of profit as 'earned' or 'unearned' income did not interest Ricardo: he devoted one sentence to this subject in the *Principles* and even this sentence was a throw-away remark.

Marx also learned from Ricardo how to reduce skilled labour to common labour by simply taking the structure of relatives wages as given, thus missing the thrust of Smith's theory of relative wages, namely, that wages are not determined solely by the demand side in labour markets.

Marx discarded the Malthusian theory of population but retained the subsistence theory of wages relying on the 'reserve army' of the unemployed to keep wages fluctuating around subsistence levels. He failed to notice, however, that this made wages a function of the play of demand and supply in labour markets and not the labour-costs of producing wage goods; in short, the pricing of wage goods in Marx does not conform to the labour theory of value. Like Ricardo, Marx conceded that the level of 'subsistence' is itself historically conditioned: it is a standard of living that workers have become accustomed to expect by past experience. Thus, even the 'natural' price of labour in Marx is not entirely cost-determined but depends on the preferences of workers. Once again, the 'value of labour-power' in Marx does not conform to the labour theory of value.

Marx never paid much attention to Ricardo's doctrine of comparative advantage and apparently failed to notice that it too violates the labour theory of value. It is also doubtful whether he ever truly grasped the import of Ricardo's theory of differential rent and particularly its central implication that prices everywhere, and not just in agriculture, are determined by marginal rather than average costs of production.

Nevertheless, despite all the obvious differences between Smith and Ricardo on the one hand and Ricardo and Marx on the other in both analytical constructs and social vision, there are so many striking similarities between them that Marxian economics is simply unimaginable without Smith, Ricardo and (although Marx did not like to admit it) John Stuart Mill. Marx went further than any of them in his grasp of business cycles, his treatment of technical change and the so-called 'reproduction schema' – the true starting point of the modern theory of steady-state growth – but he never emancipated himself from his starting point in classical economics with all its strengths and all its weaknesses.

There can be little quarrel, therefore, with a surplus interpretation of classical economics that treats Marx squarely as one of the last classical economists. However, it is when this Marxian strand in the surplus interpretation is combined with the Sraffian strand that we begin to encounter a mythical classical economics that never existed. We are told that the data for the analysis of prices in classical economics are the same as those for Sraffa, namely, (1) the size and composition of output, (2) the techniques of production in use, and (3) the real wage rate; these are contrasted with the data of neoclassical economics, namely, the preferences of individuals, the initial endowment of the factors of production among individuals and the existing techniques of production (e.g. Eatwell 1977, p. 62). We are even told that long-run prices in classical theory are *not* the outcome of the opposing forces of demand and supply and that classical 'natural' prices are *not* what (ever since Marshall) are called long-run 'normal' prices (Harcourt 1982, p. 265) or that, although classical 'natural' prices are indeed the same as neoclassical long-run 'normal' prices, the theories advanced by classical and neoclassical economists for the determination of these long-run equilibrium prices are quite different (Garegnani 1976, pp. 28–9). But there is actually no warrant for any of these assertions.

The size and composition of output is certainly not treated as given in Smith and to say so is to make nonsense of Smith's emphasis on secular economic development and the optimum balance of manufacturing and agriculture in the course of secular growth. Ricardo, on the other hand, frequently but not invariably treats the output of agricultural produce as determined by the size of population via a perfectly inelastic demand for wheat (Barkai 1965; Stigler 1965). Thus, he does not assume the output of wheat (or any other product) to be a datum but to be an endogeneously determined variable, a function of population growth, which in turn is treated as an endogeneous variable. He never squarely faced up to all the difficulties created for his argument by commodity-substitution as the price of 'corn' rises relative to 'cloth', but he certainly recognized the problem. There is no support, therefore, for the contention that he took the composition of output to be a datum, except provisionally at certain points in his argument for the sake of producing what he called 'strong results'. What we have said about Smith and Ricardo follows with double force for both Mill and Marx. So much then for

this part of the attempt to bring the classical economists fully into the Sraffian fold.

We can agree that the classical economists took for granted an existing state of techniques – has there ever been an economist, apart possibly from Marx, who has not? – but the real question is whether they conceived of this state of techniques *à la* Sraffa as ruling out factor substitution. On balance, as we noted earlier, the answer to this question must be yes. Ricardo of course recognized the problem the moment he introduced the chapter on machinery in the third edition of the *Principles* (1821), but by then he was thoroughly committed to his invariable standard of value, which necessarily rules out factor substitution. On the other hand, a special kind of factor substitution was built into his theory of differential rent in which variable doses of capital-and-labour combined in fixed proportions are applied in increasing amounts to a fixed quantity of heterogeneous land; it is this idea which of course led John Bates Clark and Philip Wicksteed in later years to hail Ricardo as the 'father' of marginal productivity theory. When we consider that the theory of differential rent was the very cornerstone of the Ricardian system, we can only gasp at Sraffa's bold declaration in the preface to his *Production of Commodities by Means of Commodities* (1960) that his own system, concerned as it is 'exclusively with such properties of an economic system as do not depend on changes in the analysis of value production or in the proportions of "factors" is identical to the 'standpoint . . . of the old classical economists from Adam Smith to Ricardo'.

Next, can it be argued that the classical economists took the real wage rate as a datum for their analysis of value and distribution? It is perfectly true that the much-maligned theory of subsistence wages in factor amounts to saying that the subsistence wage is whatever has been the real wage for a long time. How long is long? About a generation, Malthus said, and Ricardo agreed. But such assertions did not help much in specifying the subsistence wage, since annual population growth had been positive for as long as anyone could remember, and a positive rate of population growth implied that market wages exceed the natural subsistence wage rate. So, in effect, the classical economists regarded real wages as data but that is not what they thought they were doing; after all, the only reason that the Malthusian theory of population was so quickly incorporated into the mainstream of classical economics was that it appeared to provide a truly endogeneous explanation of the determination of real wages. The long-run equilibrium wage rate, Malthus had taught, was that wage rate, which, given the historically conditioned habits and customs of the working class, encouraged them to reproduce a family of given size. Some classical economists, like Senior and McCulloch, came to doubt the validity of the Malthusian theory but never managed to put any other theory of determination of long-run wages in its place. John Stuart Mill, on the other hand, found the Malthusian theory so suitable for his purpose of alleviating poverty through the self-help of the poor – birth control, education and the formation of consumer and producer cooperatives – that he espoused it more vehemently than even Malthus himself. All in all, there is simply no warrant for arguing that any classical economist (including Marx) *intended* to explain real wages by forces outside the purview of economic analysis.

Lastly, we come to the most grotesque distortion of all: the idea that any appeal to the forces of demand *and* supply in determining prices is necessarily alien to classical economics and that classical 'natural' prices have nothing whatsoever in common with Marshall's long-run 'normal' prices. Now, it is true that Ricardo (and Marx after him) propagated the misleading idea that demand-and-supply explanations only pertain to 'market' prices, whereas 'natural' prices are to be explained solely in terms of costs of production, as if costs can influence prices without acting through supply. Ricardo lacked the analytical apparatus to appreciate the fact that supply-side explanations of prices hold only if goods are produced under conditions of constant costs; this might well justify the neglect of demand in the case of the pricing of 'cloth' but certainly not on his own grounds in the case of the pricing of 'corn'. This marvellous confusion of language, encouraged by Ricardo's tendency to think of

demand and supply as quantities actually bought and sold and not as schedules of demand and supply prices, was almost entirely cleared up by Mill in his masterful treatment of value in Book III of his *Principles* in which he noted that an equilibrium price is one which equates demand and supply in the sense of a mathematical equation and concluded that 'the law of demand and supply . . . is controlled but not set aside by the law of cost of production, since cost of production would have no effect on value if it could have none on supply'. In fact, this is not very different from what Ricardo (1952, Vol. IX, p. 172) once said in private to Jean Baptiste Say: 'You say demand and supply regulates the price of bread; that is true, but what regulates supply? The cost of production.'

Marshall's schema of market-period, short-period and long-period prices, of constant-cost, increasing-cost and decreasing-cost industries, and their accompanying diagrams of demand and supply, are indispensable aids to clear thinking about the determination of prices and imply nothing whatsoever about the truth or falsity of any particular theory of prices. To treat demand and supply as dirty words that classical economists would never have employed in the explanation of natural prices is to take their outmoded language at its face value and, indeed, to deny any analytical progress in the history of economics.

To reject Sraffian interpretations of classical economics is not to reject Sraffa's system on its own grounds. Whether or not it is faithful to both the spirit and the letter of classical economics, it is undeniably true that, like all advances in economic theory, it casts a new light on the ideas of the past. It has certainly made us think again about Ricardo's invariable measure of value and its intimate connection with Marx's transformation problem; it has illuminated the problem of joint production and the difficulties which this creates for the labour theory of value, however formulated; and it has highlighted the fact that any theory of prices necessarily involves some proposition about how total output is divided between wages and profits. Its impact on the ongoing debate about the great ideas of the past is perhaps best illustrated by the furore which it has created among Marxian economists, suggesting for example, that the labour theory of value in Marx is both unnecessary and incapable of producing Marx's results (Steedman 1977, 1981). But to endorse Sraffa's system as a tool for historical exegesis is not to say that it successfully models the essence of classical economics. Smith, Ricardo, Mill and Marx are simply richer than anything captured in *Production of Commodities by Means of Commodities.*

Classical Economics as General Equilibrium Theory

Every extreme reaction produces a counter-reaction. The surplus interpretation of classical economics is a reaction against Marshallian interpretation of classical economics in which Ricardo and Mill are viewed as neoclassical theorists in embryo; for Marshall there was one and only one thread of continuous thought from Adam Smith to his own times (e.g. Marshall 1890, App. I). In reaction to the surplus interpretation, Hollander has argued that from Ricardo onwards, classical economics was, for all practical purposes, general equilibrium theory; there never was any 'marginal revolution'. Since this assertion is, to say the least, surprising, let us quote his own words:

> Ricardian economics – the economics of Ricardo and J.S. Mill – in fact comprises in its essentials an exchange system fully consistent with the marginalist elaborations. In particular, their cost–price analysis is pre-eminently an analysis of the allocation of scarce resources, proceeding in terms of general equilibrium, with allowance for final demand, and the interdependence of factor and commodity markets. (Hollander 1982, p. 590.

It is evident that by 'general equilibrium theory', Hollander means a number of interconnected propositions, such as efficient allocation of given resources among alternative uses subject to the principle of diminishing marginal returns, the simultaneous determination of both quantities and relative prices with the aid of the principle of equality between demand and supply, and the consequent interdependence between equilibrium in product and factor markets. Perhaps we have

already said enough to suggest that if this what is meant by general equilibrium theory, there is no sense in which we can subscribe to Hollander's interpretation of classical economics.

Hollander has spelled out his meaning in great detail in a major work on *The Economics of David Ricardo* (1979). In interpreting Ricardo as a general equilibrium theorist, Hollander found himself revising more or less the entire body of Ricardian scholarship, implying that absolutely everybody else before him had radically misinterpreted Ricardo. To convey the flavour of his iconoclasm, consider the following small sample of the extraordinary conclusions of this book (for a complete list, see O'Brien 1981, pp. 354–5): (1) Ricardo's method of analysis was identical to that of Adam Smith; (2) Ricardo's theory of money was not very different from that of Smith; (3) Ricardo treated the pricing or products and the pricing of factors as fully interdependent; (4) Ricardo's profit theory did not originate in a concern over the Corn Laws, and Ricardo never believed, even in his early writings, that profits in agriculture determine the general rate of profit in the economy; (5) Ricardo's value theory was essentially the same as that of Marshall in that it paid as much attention to demand as to supply, and Ricardo never regarded the invariable measure of value as an important element in his theory; (6) Ricardo could have established his fundamental theorem of the inverse wage–profit relationship without his invariable yardstick and he frequently took the short-cut of assuming identical capital–labour ratios in all industries to give the answers he looked for; (7) wages in Ricardo are never conceived at any time as constant or fixed at subsistence levels; (8) Ricardo never assumed a zero price-elasticity of demand for corn, making the demand for agricultural produce a simple function of the size of population; (9) Ricardo did not predict a falling rate of profit or a rising rental share and never committed himself to any clear-cut predictions about any economic variable; and (10) Ricardo was never seriously concerned about the possibility of class conflict between landowners and everybody else or between workers and capitalists.

There must be something wrong with an interpretation of Ricardo that produces so many conclusions diametrically opposed to what every commentator has found in Ricardo, not only since his death but even while he was still alive. The distortions produced by the surplus interpretation of classical economics are therefore as nothing compared to those generated by Hollander's general equilibrium interpretation.

Walsh and Gram (1980) provide a more reasonable version of the general equilibrium characterization of classical economics: they take the view that general equilibrium analysis encompasses more or less the whole of the history of economic thought, but they distinguish between pre-Walrasian general equilibrium analysis of the allocation of the economic surplus over successive time periods and post-Walrasian general equilibrium analysis of the allocation of given resources within the same time period. One difficulty with their argument is that they never inform the reader what precisely is meant by 'general equilibrium analysis'. If we mean a discussion of the determination of both product and factor prices which proceeds in terms of an explicit or implicit set of simultaneous equations in order to ensure that the number of unknowns to be determined are equal to the number of equations written down, then obviously classical economics is not general equilibrium analysis: factor pricing in classical economics is invariably explained on different principles from those governing the pricing of products. If we go further and demand that such a discussion must include not just a demonstration of the existence of a unique equilibrium solution for the vector of factor and product prices but also an analysis of the stability and determinacy of the set of equilibrium prices, such as Walras himself struggled to provide, then even more obviously classical economics is not general equilibrium analysis. But what Walsh and Gram seem to mean by general equilibrium analysis is simply any analysis that involves the simultaneous determination of prices and one distribution variable on the assumption that other factor prices are given; in short, they define general equilibrium analysis to be nothing more nor less than Sraffian economics. Their book therefore collapses the

general equilibrium interpretation of classical economics into the surplus interpretation, sharing the deficiencies of both in equal proportions.

Finally, Arrow and Hahn (1971, pp. 1–3) join the fray in the introduction to their textbook on general equilibrium theory. In contrast to Walsh and Gram, they are perfectly explicit about what is meant by general equilibrium theory: if it means anything it implies some notion of both determinateness and stability, that is, the relations describing the economic system are sufficient to determine the equilibrium values of its variables, and a violation of any one of these relations sets in motion forces to restore it. They go on to introduce a new note into the argument: general equilibrium theory is typically associated with the doctrine of unintended consequences – equilibrium outcomes may be and usually are different from those intended by individual actors – and the doctrine that competition is a social mechanism that is capable of achieving a determinate and stable set of equilibrium prices. In all these senses of the term, they count Adam Smith as a 'creator' of general equilibrium theory and Ricardo, Mill and Marx as early expositors. They add, however, that there is another sense in which none of the classical economists had a 'true general equilibrium theory': no classical economist gave explicit attention to demand as a coordinate element with supply in determining prices, and hence classical economics determined the prices but not the quantities of commodities, the only exception to this statement being their treatment of agricultural output; on the other hand, Mill's theory of foreign trade was 'a genuine general equilibrium theory'.

To this brief but incisive discussion of the sense in which classical economics is or is not general equilibrium theory, one must add one word of caution: it is the subtle but nevertheless unmistakable difference in the conception of 'competition' before and after the 'marginal revolution'. The modern concept of perfect competition, conceived as a market structure in which all producers are price-takers and face perfectly elastic sales curves for their outputs, was born with Cournot in 1838 and is foreign to the classical conception of competition as a process of rivalry in the search for unrealized profit opportunities, whose outcome is uniformity in both the rate of return on capital invested and the prices of identical goods and services but not because producers are incapable of making prices. In other words, despite a steady tendency throughout the history of economic thought to place the accent on the end-state of competitive equilibrium rather than the process of disequilibrium adjustments leading up to it, this emphasis became remorseless after 1870 or thereabouts, whereas the much looser conception of 'free competition' with free but not instantaneous entry to industries is in evidence in the work of Smith, Ricardo, Mill, Marx and of course Marshall and modern Austrians (Stigler 1957; McNulty 1967; Littlechild 1982). For that reason, if for no other, it can be misleading to label classical economics as a species of general equilibrium theory except in the innocuous sense of an awareness that 'everything depends on everything else'.

Summing Up

We have reviewed the recent upswell of new and startling interpretations of classical economics in the light of developments in modern economics, such as the economics of development, growth theory, general equilibrium theory, and Sraffian analysis. In itself there is nothing surprising about this, nor is it a new phenomenon: every turn and twist in the history of economic thought has always been attended by a fresh look at the past. Marx in propounding his own treatment of the 'laws of motion' of capitalism felt impelled to re-examine the ideas of his predecessors over more than a thousand pages. Jevons, Menger and Walras, the triumvirate that is said to have launched the 'marginal revolution', accompanied the exposition of their 'new' economics by scathing denunciations of the fallacies of classical political economy. Marshall, in seeking unsuccessfully to reconcile a static with a dynamic treatment of economic problems, naturally looked with sympathy at the work of his classical forebears and struggled to depict them as slightly exaggerating one side of the truth in contrast to Jevons, who exaggerated the other. Perhaps

therefore the recent proliferation of definitely new but conflicting interpretations of the essential meaning of classical economics is simply an expression of the fact that modern economists are divided in their views and hence quite naturally seek comfort by finding (or pretending that they can find) these same views embodied in the writings of the past.

See Also

- ▶ Classical Growth Models
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Mill, John Stuart (1806–1873)
- ▶ Ricardo, David (1772–1823)
- ▶ Smith, Adam (1723–1790)

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.

Barkai, H. 1965. Ricardo's static equilibrium. *Economica* 32: 15–31.

Blaug, M. 1958. *Ricardian economics: An historical study.* New Haven: Yale University Press.

Blaug, M. 1985. *Economic theory in retrospect*. 4th ed. Cambridge: Cambridge University Press.

Bloomfield, A.I. 1975. Adam Smith and the theory of international trade. In *Essays on Adam Smith*, ed. Skinner A. and Wilson T. Oxford.

Bloomfield, A.I. 1978. The impact of growth and technology of trade in nineteenth-century economic thought. *History of Political Economy* 10: 608–635.

Bloomfield, A.I. 1981. British thought on the influence of foreign trade and investment of growth, 1800–1880. *History of Political Economy* 13: 95–120.

Bradley, I., and M.C. Howard. 1982. *Classical and Marxian political economy.* London: Macmillan.

Buchanan, D. 1929. The historical approach to rent and price theory. *Economica* 9 (26): 123–155.

Cropsey, J. 1957. *Polity and economy: An interpretation of the principles of Adam Smith*. The Hague: Nijhoff.

Dobb, M. 1973. *Theories of value and distribution since Adam Smith*. London: Cambridge University Press.

Eagly, R.V. 1974. *The structure of classical economic theory*. New York: Oxford University Press.

Eatwell, J. 1977. The irrelevance of returns to scale in Sraffa's analysis. *Journal of Economic Literature* 15 (1): 61–68.

Eatwell, J. 1982. Competition. In *Classical and Marxian political economy*, ed. I. Bradley and M. Howard. London: MacMillan.

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

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

Garegnani, P. 1984. Value and distribution in the classical economists and Marx. *Oxford Economic Papers* 36: 291–325.

Gordon, D.F. 1959. What was the labour theory of value? *American Economic Review* 49: 462–472.

Harcourt, G.C. 1982. The Sraffian contribution: An evaluation. In *Classical and Marxian political economy*, ed. I. Bradley and M. Howard. London: MacMillan.

Hollander, S. 1973. *The economics of Adam Smith.* Toronto: University of Toronto Press\London: Heinemann Educational Books.

Hollander, S. 1977. The reception of Ricardian economics. *Oxford Economic Papers* 29: 221–257.

Hollander, S. 1979. *The economics of David Ricardo.* Toronto: University of Toronto Press\London: Heinemann Educational Books.

Hollander, S. 1981. Marxian economics as 'general equilibrium' theory. *History of Political Economy* 13: 121–155.

Hollander, S. 1982. On the substantive identity of the Ricardian and neoclassical conception of economic organization: The French connection in British classicism. *Canadian Journal of Economics* 15: 586–612.

Howard, M.C., and J.E. King. 1985. *The political economy of Marx*. 2nd ed. London: Longman.

Kaldor, N. 1972. The irrelevance of equilibrium economics. *Economic Journal* 82 (December): 1237–1255.

Keynes, J.M. 1936. The general theory of employment, interest and money. In *The collected writings of John Maynard Keynes*, vol. 7. London: Macmillan, 1973.

Knight, F.H. 1956. *On the history and method of economics*. Chicago: University of Chicago Press.

Laidler, D.E.W. 1981. Adam Smith as a monetary economist. *Canadian Journal of Economics* 14 (2): 187–200.

Lindgren, J.R. 1973. *The social philosophy of Adam Smith*. The Hague: Martinus Nijhoff.

Littlechild, S.C. 1982. Equilibrium and the market process. In *Method, process, and Austrian economics*, ed. I.M. Kirzner. Lexington: D.C. Heath.

Lowe, A. 1954. The classical theory of economic growth. *Social Research* 21 (Summer): 127–158.

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

Marx, K. 1867. *Capital: A critique of political economy.* Trans. B. Fowkes. Harmondsworth: Penguin Books, 1976.

McNulty, P.J. 1967. A note on the history of perfect competition. *Journal of Political Economy* 75: 395–399.

Meek, R.L. 1967. *Economics and ideology and other essays*. London: Chapman & Hall.

Meek, R.L. 1973. *Studies in the labour theory of value*. 2nd ed. London: Lawrence & Wishart.

Meek, R.L. 1977. *Smith, Marx and after.* London: Chapman & Hall.

Mill, J.S. 1844. *Essays on some unsettled questions political economy*. London: London School of Economics, 1948.
Myint, H. 1948. *Theories of welfare economics*. Cambridge, MA: Harvard University Press.
Myint, H. 1958. The 'classical theory' of international trade and underdeveloped countries. *Economic Journal* 68: 317–337.
O'Brien, D.P. 1975. *The classical economists*. London: Oxford University Press.
O'Brien, D.P. 1981. Ricardian economics and the economics of David Ricardo. *Oxford Economic Papers* 33: 352–386.
Ong, N.-P. 1983. Ricardo's invariable measure of value and Sraffa's 'standard commodity'. *History of Political Economy* 15: 207–227.
Pasinetti, L.L. 1974. *Growth and income distribution: Essays in economic theory*. Cambridge: Cambridge University Press.
Peach, T. 1984. David Ricardo's early treatment of profitability: A new interpretation. *Economic Journal* 94: 733–751.
Peacock, A. 1975. The treatment of the principles of public finance in *The Wealth of Nations.* In *Essays on Adam Smith*, ed. Skinner, A. S. and Wilson, T. Oxford.
Ricardo, D. 1951–73. *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press.
Robertson, H.M., and W.L. Taylor. 1957. Adam Smith's approach to the theory of value. *Economic Journal* 67 (June): 181–198.
Roncaglia, A. 1978. *Sraffa and the theory of prices*. New York: Wiley.
Rosenberg, N. 1960. Some institutional aspects of the *Wealth of Nations*. *Journal of Political Economy* 68: 557–570.
Samuelson, P. 1978. The canonical classical model of political economy. *Journal of Economic Literature* 16: 1415–1434.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Skinner, A.S. 1979. *A system of social science: Papers relating to Adam Smith*. Oxford: Clarendon Press.
Skinner, A.S., and T. Wilson. 1975. *Essays on Adam Smith*. Oxford: Clarendon Press.
Sowell, T. 1974. *Classical economics reconsidered*. Princeton: Princeton University Press.
Spiegel, H.W. 1983. *The growth of economic thought*. Durham: Duke University Press.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Steedman, I. 1977. *Marx after Sraffa*. London: New Left Books.
Steedman, I., ed. 1981. *The value controversy*. London: Verso Editions and New Left Books.
Stigler, G.J. 1957. Perfect competition, historically contemplated. *Journal of Political Economy* 65: 1–17.
Stigler, G.J. 1965. Textual exegesis as a scientific problem. *Economica* 33: 447–450.
Thweatt, W.O. 1976. James Mill and the early development of comparative advantage. *History of Political Economy* 8: 207–234.
Walsh, V., and H. Gram. 1980. *Classical and neoclassical theories of general equilibrium*. Oxford: Oxford University Press.
West, E.G. 1976. Adam Smith's economics of politics. *History of Political Economy* 8: 515–539.
White, L.H. 1984. *Free banking in Britain: Theory, experience and debate, 1800–1845*. Cambridge: Cambridge University Press.
Winch, D. 1978. *Adam Smith's politics: An essay in historiographic revision*. Cambridge: Cambridge University Press.
Winch, D. 1983. Science and the legislator: Adam Smith and after. *Economic Journal* 93: 501–520.

Bronfenbrenner, Martin (Born 1914)

E. R. Weintraub

Bronfenbrenner received his AB from Washington University (St Louis) and his Ph.D. from the University of Chicago in 1939. A student of Oskar Lange, Bronfenbrenner's professional career and writing reflect a catholicity of interests rare among modern economists. His books range from the careful and judicious treatise *Income Distribution Theory* (1971) to the playful collection of short stories on the American occupation of Japan, *Tomioko Stories* (1976). He is probably the only person to hold simultaneous memberships in the conservative Mt Pelerin Society and the Union of Radical Political Economists.

He has taught, and written on, macroeconomics, trade theory and policy, monetary economics, production theory, development economics, the history of economic thought, distribution theory, economic history, Marxian economics, and the Japanese economy. He is one of the most prolific of contemporary economists, his writings being characterized by elegance and felicitous phrasing and further adorned by verses from obscure poets and popular operettas.

As a 'neoclassical' economist, trained at Chicago, his contributions to economic analysis themselves blend a variety of themes and techniques. His major work on income distribution theory itself modifies neoclassical theory so that it can frame

questions raised in both classical and neo-Marxian analysis. He has been a leader in the analysis of Japanese economic development and growth, an interest fostered during his military service as a Japanese language officer, and as an economist attached to the US occupation forces in Japan.

His regular academic appointments included positions at Wisconsin, Michigan State, Minnesota, Carnegie Tech, Aoyoma Gakuin and Duke (where he was Kenan Professor of Economics). Many of his professional papers are held by the Manuscript Department of Perkins Library, Duke University.

Selected Works

1945. Some fundamentals of liquidity theory. *Quarterly Journal of Economics* 59, May, 405–426.
1956. Potential monopsony in labour markets. *Industrial and Labour Relations Review* 9, July, 577–588.
1959 (ed.) *Is the business cycle obsolete?* New York: Wiley.
1961. *Academic encounter*. New York: Free Press.
1961. Some lessons of Japanese economic development, 1868–1938. *Pacific Affairs* 34(1), Spring, 7–27.
1963. (With F.D. Holzman.) Survey of inflation theory. *American Economic Review* 53, September, 593–661.
1971. *Income distribution theory*. Chicago: Aldine.
1976. *Tomioko stories*. New York: Exposition Press.
1979. *Macroeconomic Alternatives*. Chicago: AHM Publishing Co.
1984. (With W. Sichel and W. Gardner.) *Economics*. Boston: Houghton Mifflin.

Brougham, Henry (1773–1868)

F. Y. Edgeworth

Baron Brougham and Vaux, Lord Chancellor, touched nearly all subjects and adorned some by his eloquence and dialectical skill. The contact seems least superficial, the ornament particularly solid, in the case of political economy. Brougham's first considerable work was *An Inquiry into the Colonial Policy of European Powers*, 1803. Criticizing Adam Smith, he maintains that the monopoly of the colonial trade did not produce all the detrimental effects ascribed to it (Book I, §2, part ii). Referring to the slave colonies, Brougham not only denounces the slave trade as iniquitous – 'not a trade, but a crime' – but also argues that it is unprofitable. The argument is renewed in *A Concise Statement of the Question regarding the Abolition of Slave Trade* (1804). Slavery, as well as slave trade, was assailed by Brougham's oratory (*Speeches*, published in 1838, vol. ii.).

Free trade owes something to Brougham's advocacy. He exposed the folly of retaliation, as counsel (1808) for the merchants who petitioned parliament against the orders in council directed against Napoleon's continental system. After Brougham's masterly speech in 1812, the obvious orders were withdrawn (*Speeches*, vol. i). In the speech on manufacturing distress (1817) Brougham strikes at the complicated taxes which fettered trade (ibid.) But in the equally able speech on agricultural distress (1816) there is a good word for the corn law (ibid. p. 533).

Other economic topics handled by Brougham are: (1) depreciation of money, with reference to Sir E. Shuckburgh's standard (article on 'Currency and Commerce' [1803], *Contributions to Edinburgh Review*, published 1856, vol. iii, p. 22); (2) usury ([1816], *Contributions*, vol. iii p. 52); (3) over-population (speech on the Poor Laws, 1834, *Speeches*, vol. iii); (4) combinations (*Transactions of the Society for promoting Social Science* for 1860, p. 51). Brougham is also to be mentioned as a promoter of education and educational institutions – the London University, the Society for the Diffusion of Useful Knowledge, Mechanics' Institutes, and the Society for promoting Social Science.

In addition to the works which have been cited may be noticed: (1) *A Manual for Mechanics' Institutions*, 1839 (by B.F. Duppa, with outlines of lectures on political economy by Brougham); (2) *Political Philosophy*, 1842; (3) *Works*, 1st edn

1855–61, 2nd edn 1873. In the 11th volume of the 2nd edition there is a list of Brougham's publications, numbering 133.

Selected Works

1803. *An inquiry into the colonial policy of European powers*, 2 vols. Edinburgh: E. Balfour.
1804. *A concise statement of the question regarding the abolition of slave trade*. London: J. Hatchard and T.N. Longman.
1838. *Speeches of Henry Lord Brougham*, 4 vols. Edinburgh: A. & C. Black.
1839. Outlines of lectures on political economy. In *A manual for mechanic's institutions,* ed. B.F. Duppa. London: Society for the Diffusion of Useful Knowledge.
1842. *Political philosophy*. London: Society for the Diffusion of Useful Knowledge.
1855–61. *Works*, 11 vols. London, Glasgow: R. Griffin & Co.; 2nd ed, Edinburgh: A. & C. Black, 1873.
1856. *Contributions to the Edinburgh review*, 3 vols. London, Glasgow: E. Griffin & Co.

Brown, Harry Gunnison (1880–1975)

Mason Gaffney

Harry G. Brown was born in Troy, New York, the son of an accountant. He was stricken from age four with tuberculosis of the hip. He graduated from Williams (1904) and took his PhD at Yale in 1909.

Brown instructed at Yale from 1909 to 1915, working closely with Irving Fisher. Herbert J. Davenport then hired him at the University of Missouri where he succeeded Davenport as Chair, remaining there through 1947. After retirement in 1950 he taught at The New School, The University of Mississippi, and Franklin and Marshall. During the Pennsylvania residence he campaigned vigorously, although in his eighties, for adoption of the graded tax in cities of that state, under its local-option law. He returned to Columbia, Missouri to retire, and continued publishing and speaking until his death at the age of 95.

Brown was the premier exponent among orthodox theorists of taxing land values and saw rail/utilities as generating taxable rents. There is producer surplus when rates exceed average cost. There is also consumer surplus when rates fall below average benefits. Both surpluses were locational and lodged in land rents. This led to a general interest in taxing land. He used conventional tools, careful craftsmanship, and a priori methods. Brown refuted J.B. Clark's idea that 'the lure of unearned increment' was a constructive incentive for pioneering; and Frank Knight's idea that land is like all other resources because it has an opportunity cost; and Ely's doctrine that 'ripening costs' of land speculators justify urban sprawl.

He criticized most non-land taxes, but opposed Federal deficits more than any tax, seeing deficits as competing directly for capital formation. He attributed macro problems to cost–push, pointing to overpricing of land and raw materials, which land taxation would abate; to high interest rates forced by inadequate bank reserves; to associationism in industry; and to the perverse behaviour of regulated rates under average-cost pricing, where reduced demand forces rates to rise and worsen a depression. He would increase investment opportunities by abating taxes on capital.

Selected Works

1931. *Economic science and the common welfare*. Columbia: Lucas Bros.
1980. In *The selected articles of Harry Gunnison Brown*, ed. P. Junk. New York: The Robert Schalkenbach Foundation.

References

Essays in Honor of Harry Gunnison Brown. *American Journal of Economics and Sociology* 11(3), 1952.

Brunner, Karl (1916–1989)

Allan H. Meltzer

Keywords

Aggregate demand; Brunner, K.; Information costs; Interest rate; Macroeconomic theory; Monetarism; Monetary theory; Money; Money supply; Patinkin, D.; Shocks; Stocks and flows; Transaction costs; Uncertainty

JEL Classifications

B31

Karl Brunner's scholarly contributions are in three areas, namely, monetary and macroeconomics, methodology and its application to cognitive science, and social, political, and institutional analysis. Brunner founded three major journals and organized many conferences, including the Konstanz Seminar in Germany and the Carnegie-Rochester Conference in the United States, which remain current in 2007. Laidler (1991) contains a more complete discussion of Brunner's contributions, and I have relied heavily on his paper. Brunner's own discussion of his intellectual and personal odyssey is in Brunner (1988). I was involved as co-author in much of the work on monetary economics, but I choose to use the pronouns 'he' and 'his' for this article.

Brunner was born in Zurich, Switzerland, in February 1916. His mother was from the French-speaking region, his father from the German-speaking. They met when both were in Russia working with Russian children. Later his father became the director of the Swiss Observatory. Karl received his doctorate in economics from the University of Zurich in 1943 after spending 1937–38 studying modern economics at the London School of Economics. He travelled to the United States as a Rockefeller Foundation Scholar at Harvard and the University of Chicago from 1949 to 1951. He served on the UCLA faculty from 1951 to 1966 when he left on visiting appointments at Wisconsin and Michigan State before becoming the Everett D. Reese Professor of Economics at Ohio State University. In 1966, he moved to the University of Rochester, where he remained until his death in 1989. From 1979 to 1989, he was the Fred H. Gowen Professor of Economics. During his years at Rochester he served also as Permanent Guest Professor at the University of Konstanz (Germany) from 1968 to 1973 and Professor Ordinarius at the University of Bern (Switzerland) from 1974 to 1985. He arranged for many of his doctoral students at Bern to study at the University of Rochester. This had a lasting influence on economics and finance in Switzerland and Europe.

Brunner often commented on the gap, often a wide one, between economic policy and economic theory. Much of his research, his efforts to influence policy, his journals and conferences reflected his belief that this gap could be closed by substantive research. Much of his analysis of institutions and the policy process considered the incentives that produced these outcomes and the uncertainty under which policies are made. To properly analyse issues of this kind, he proposed (1987) replacing the 'economic man' of the textbooks with the more dynamic and uncertain REMM – resourceful, evaluating, maximizing man. He used REMM also to compare economists', sociologists', political scientists' and psychologists' ability to understand society's processes.

Macroeconomic theory and monetary theory were his major interests. His earliest work (1951) was a lasting contribution to the early post-war concern with the purely analytic issues raised by Don Patinkin and others as to the determinacy of equilibrium in classical macroeconomics. Brunner developed a stock–flow analysis and devised equilibrium conditions.

Purely formal analysis did not fit well with his developing ideas about methods and the means to scientific development and knowledge in economics. He saw economics as an empirical science that produced refutable hypotheses. He did not reject formal analysis; he no longer did it.

After a few years, he turned to money supply theory. The central idea was to go beyond the standard IS–LM framework in which typically bonds and real capital are perfect substitutes, so

that a single interest rate could represent the panoply of relative prices that transmit monetary and other impulses through the economic system. Brunner began by making the interest rate and the money supply endogenous variables. This generation of models was used to reject reverse causation and to critique Federal Reserve policymaking in a study for the US Congress. He proposed an alternative (Brunner and Meltzer 1964).

Subsequent work (Brunner and Meltzer 1989) introduced an output sector with endogenous prices and output. The complete static model had two endogenous relative prices, base money, bonds and real capital. Adding some institutional detail brought in the money stock and bank credit.

Although anticipated prices appear in these models, price expectations have a minimal role. Responding to the heightened emphasis in the 1970s on expectations and many discussions of stagflation, Brunner et al. (1980, 1983) introduced transitory and persistent shocks into the analysis. This offered an explanation of asset markets requiring at least two relative prices to account for uncertainty of beliefs about the persistence of various impulses. It also offered an explanation of gradual adjustment of wages and employment in response to shocks of uncertain duration. The extended (1983) model introduced price setting and allowed inventories to absorb short-run shocks to aggregate demand.

The role of uncertainty and information was recognized early but took a central position in his monetary theory in 'The Uses of Money' (Brunner and Meltzer 1971). The paper develops the reason that society adopts money, treats money's central role as a medium of exchange and explains why societies converge to a small number, often a single, money. The medium of exchange reduces transaction and information costs, thereby saving resources.

Karl Brunner is known as one of the founders of monetarism, a name he coined for the counter-revolution against Keynesian economics of the 1950s and 1960s.

See Also

► Monetarism

Selected Works

1951. Inconsistency and indeterminacy in classical economics. *Econometrica* 19: 152–173.
1964. (With A.H. Meltzer.) *The federal reserve's attachment to the free reserve concept.* Washington: House Committee on Banking and Currency.
1971. (With A.H. Meltzer.) The uses of money: Money in the theory of an exchange economy. *American Economic Review* 71: 784–805.
1980. (With A. Cukierman and A.H. Meltzer.) Stagflation, persistent unemployment, and the permanence of economic shocks. *Journal of Monetary Economics* 6: 467–492.
1983. (With A. Cukierman and A.H. Meltzer.) Money and economic activity, inventories and business cycles. *Journal of Monetary Economics* 11: 281–319.
1987. The perception of man and the conception of society: Two approaches to understanding society. *Economic Inquiry* 25: 367–88.
1988. My philosophy. *American Economist.*
1989. (With A.H. Meltzer.) *Money and the economy: Issues in monetary analysis.* The 1987 Raffaele Mattioli Lectures. Cambridge: Cambridge University Press.

Bibliography

Laidler, D. 1991. Karl Brunner's monetary economics: An appreciation. *Journal of Money, Credit, and Banking* 23: 633–658.

Bruno, Michael (1932–1996)

Joseph Zeira

Keywords

Bruno, M.; Falk Institute for Economic Research (Israel); Inflation; Intertemporal approach to the balance of payments; Oil and the macroeconomy; Sachs, J.; Stabilization policy; Stagflation; World Bank

JEL Classifications
B31

Michael Bruno was born in Germany in 1932 and emigrated with his family to Israel in 1933. After military service he studied mathematics and economics at the Hebrew University of Jerusalem and at King's College, Cambridge. On returning to Israel, he worked at the research department of the Bank of Israel. In 1961 he was brought to Stanford University by Hollis Chenery and Kenneth Arrow, where he received his Ph.D. in 1962. He then returned to Israel and in 1963 joined the faculty of the Department of Economics at the Hebrew University of Jerusalem. Over the years he visited MIT, Harvard, the University of Stockholm, and the LSE. Many times during his academic career Michael Bruno was involved in economic policymaking. In the mid-1970s he participated in a tax reform in Israel and advised the government on economic policy. In 1985 he was chief advisor to the Israeli disinflation programme. From 1986 to 1991 he was Governor of the Bank of Israel, and between 1993 and 1996 he served as a Senior Vice-President and Chief Economist of the World Bank.

Michael Bruno's research covered many areas in macroeconomics, was both theoretical and empirical, but was always strongly related to the economic problems of the time. In the 1960s, living in a rapidly developing country, he studied economic growth and development, focusing on input–output analysis and on duality in growth theory. In the 1970s, following the oil shocks, he began to study the macroeconomics of open economies, especially their reaction to shocks. One outcome of this research contains a pioneering discussion of the important 'intertemporal approach to the balance of payments' (1976). Another outcome is the research conducted with Jeffrey Sachs on stagflation and supply shocks, which culminated in their important book on stagflation (1985). In the 1980s, influenced by high inflation in Israel and by his role in the Israeli stabilization programme of 1985, Bruno's attention turned to inflation and stabilization. His research then reflected his deep interest in issues of disinflation and of reform in general. He promoted the idea that creating consensus is important for the success of reforms, and applied it also to the analysis of the post-Communist transition in eastern Europe. In the 1990s Michael Bruno served in the World Bank and there his focus returned to issues of development, which he had studied in the beginning of his career. Actually, he combined it with his deep understanding of inflation and studied how inflation affects economic growth. His main finding appears in a paper with Easterly (1998) that shows that high inflation has a strong negative effect on growth. Thus, his last period of life and of economic res earch saw a closing of a circle, where he synthesized knowledge that he had accumulated throughout his scientific career, to analyse this important issue.

In addition to his general research and to his effect on policymaking, Michael Bruno also contributed significantly to research on the Israeli economy, both through his research and through his roles as director of the research department in the Bank of Israel, as director of the Falk Institute for Economic Research in Israel, and as Governor of the Bank of Israel.

See Also

- ▶ Factor Price Frontier
- ▶ Growth Models, Multisector
- ▶ Inflation
- ▶ Oil and the Macroeconomy

Selected Works

1962. *Interdependence, resource use and structural change in Israel*. Jerusalem: Bank of Israel.

1962. (With H. Chenery.) Development alternatives in an open economy: The case of Israel. *Economic Journal* 72: 79–103.

1966. (With E. Burmeister and E. Sheshinski.) The nature and significance of the reswitching of techniques. *Quarterly Journal of Economics* 80: 526–553.
1969. Fundamental duality relations in the pure theory of capital and growth. *Review of Economic Studies* 36: 49–62.
1976. The two-sector open economy and the real exchange rate. *American Economic Review* 66: 566–577.
1977. (With Y. Ben-Porath.) The political economy of a tax reform. *Journal of Public Economics* 7: 285–307.
1980. Import prices and stagflation in the industrial countries: A cross-section analysis. *Economic Journal* 90: 479–92.
1984. Raw materials, profits and the productivity slowdown. *Quarterly Journal of Economics* 99: 1–29.
1985. (With J. Sachs.) *Economics of worldwide stagflation*. Cambridge, MA: Harvard University Press.
1986. External shocks and domestic response: Israel's macroeconomic performance, 1965–1982. In The *Israeli economy: Maturing through crisis*, ed. Y. Ben-Porath. Cambridge, MA: Harvard University Press.
1986. Sharp disinflation strategy: Israel 1985. *Economic Policy* 1(2): 379–407.
1989. Econometrics and the design of economic reform (Presidential address). *Econometrica* 57: 275–306.
1990. (With S. Fischer.) Seigniorage, operating rules and the high inflation trap. *Quarterly Journal of Economics* 105: 353–374.
1993. *Crisis, stabilization and economic reform: Therapy by consensus. Clarendon lectures in economics*. Oxford: Clarendon Press.
1998. (With W. Easterly.) Inflation crises and long-run growth. *Journal of Monetary Economics* 41: 3–26.
1998. (With M. Ravallion and L. Squire.) Equity and growth in developing countries: Old and new perspectives on the policy issues. In *Income Distribution and High-Quality Growth*, ed. V. Tanzi. Cambridge, MA: MIT Press.

Brydges, Samuel Egerton, Bart. (1762–1837)

F. Hendriks

To anyone disposed to make a psychological study of a defunct antiquary, topographer, essayist, bibliographer, poet, novelist, and critic, and who added to these occupations the study of political economy and occasional authorship in that science, Sir Egerton Brydges would afford an excellent subject. On the good side may be placed his industry and power of research, considerable originality, and a deep acquaintance with the ancient literature of England and of foreign countries. On the bad side should be ranged his excessively morbid temperament, a craze about an assumed right to an ancient barony, an intense suspicion of the motives of those who differed from him, and an unfounded notion that he was not sufficiently rewarded for his services in the cause of learning. Much material bearing on all this exists in his *Autobiography* and *Letters from the Continent*, as well as in his voluminous published and privately printed works, which in the course of his long life extended to no less than one hundred and forty volumes. We find in a quantity of his letters, which have never been printed, addressed, from 1818 to 1832, to Mr. James S. Brooks, member of a firm of solicitors who acted for him, and with whom, in a characteristic manner, he often fell out, many striking examples of Sir Egerton Brydges' talent as a political economist. It is a curious fact that in his most desponding and brooding moments he would fly to political economy as a relaxation of thought and as a favourite study, just as many of our first-class English statesmen have relieved tension of mind and the excitement of political conflict by Homeric studies or the composition of Greek and Latin verses.

Although Sir Egerton Brydges' works contain flashes of insight into correct deductions, practical as well as theoretical, they are a good deal

disfigured by his want of study of the statistics and practice of commerce, and his ignorance of business generally.

Selected Works

1799. *Tests of the national wealth*. London.
1814. *Letters on the poor laws*. London.
1817. Reasons for a farther amendment to the Act of 54 George III *cap*. 156, being an act to amend the copyright act of 2 Anne. *Pamphleteer* 10: 493.
1818a. *A Vindication of the pending bill for the amendment of the copyright act from the misrepresentations and unjust comments of the syndics of the university library at Cambridge*. London.
1818b. *A Summary statement of the great grievance imposed on authors and publishers and the injury done to literature by the late copyright act*. London.
1818c. *Arguments for the employment of the poor*. London.
1818d. On the practicality of relieving the able-bodied poor. *Pamphleteer* 11: 133.
1819. *The population and riches of the nation considered*. Geneva: Lee Priory.
1820. *Answer to … a mode of relieving … the national debt of Great Britain*. Florence.
1821. What are riches? Geneva. Reprinted, *Pamphleteer* 20: 479.
1822. *Letter on the corn question*. London.
1834. *Autobiography, times and opinions of Egerton Brydges*. 2 vols. London.

Bubbles

Markus K. Brunnermeier

Abstract

Bubbles refer to asset prices that exceed an asset's fundamental value because current owners believe they can resell the asset at an even higher price. There are four main strands of models: (i) all investors have rational expectations and identical information, (ii) investors are asymmetrically informed and bubbles can emerge because their existence need not be commonly known, (iii) rational traders interact with behavioural traders and bubbles persist since limits to arbitrage prevent rational investors from eradicating the price impact of behavioural traders, (iv) investors hold heterogeneous beliefs, potentially due to psychological biases, and agree to disagree about the fundamental value.

Keywords

Arbitrage; Asset-pricing models; Asymmetric information; Autocorrelation; Backward induction; Bubbles; Centipede game; Central limit theorems; Co-integration; Efficient markets hypothesis; Fiat money; Gains from trade; Hedge funds; Limited liability; Noise traders; Overlapping generations model; Rational expectations; Risk aversion; Transversality condition; Unit roots

JEL Classifications

G1

Bubbles are typically associated with dramatic asset price increases followed by a collapse. Bubbles arise if the price exceeds the asset's fundamental value. This can occur if investors hold the asset because they believe that they can sell it at a higher price than some other investor even though the asset's price exceeds its fundamental value. Famous historical examples are the Dutch tulip mania (1634–7), the Mississippi Bubble (1719–20), the South Sea Bubble (1720), and the 'Roaring '20s' that preceded the 1929 crash. More recently, up to March 2000 Internet share prices (CBOE Internet Index) surged to astronomical heights before plummeting by more than 75 per cent by the end of 2000.

Since asset prices affect the real allocation of an economy, it is important to understand the circumstances under which these prices can deviate from their fundamental value. Bubbles have

long intrigued economists and led to several strands of models, empirical tests and experimental studies.

We can broadly divide the literature into four groups. The first two groups of models analyse bubbles within the rational expectations paradigm, but differ in their assumption as to whether all investors have the same information or are asymmetrically informed. A third group of models focuses on the interaction between rational and non-rational (behavioural) investors. In the final group of models traders' prior beliefs are heterogeneous, possibly due to psychological biases, and consequently they agree to disagree about the fundamental value of the asset.

Rational Bubbles Under Symmetric Information

Rational bubbles under symmetric information are studied in settings in which all agents have rational expectations and share the same information. There are several theoretical arguments that allow us to rule out rational bubbles under certain conditions. Tirole (1982) uses a general equilibrium reasoning to argue that bubbles cannot exist if it is commonly known that the initial allocation is interim Pareto efficient. A bubble would make the seller of the 'bubble asset' better off, which – due to interim Pareto efficiency of the initial allocation – has to make the buyer of the asset worse off. Hence, no individual would be willing to buy the asset. Partial equilibrium arguments alone are also useful in ruling out bubbles. Simply rearranging the definition of (net) return, $r_{t+1,s} := (p_{t+1,s} + d_{t+1,s})/p_t - 1$, where $p_{t,s}$ is the price and $d_{t,s}$ is the dividend payment at time t and state s, and taking rational expectations yields

$$p_t = E_t\left[\frac{p_{t+1} + d_{t+1}}{1 + r_{t+1}}\right] . \tag{1}$$

That is, the current price is just the discounted expected future price and dividend payment in the next period. For tractability assume that the expected return that the marginal rational trader requires in order to hold the asset is constant over time, $E_t[r_{t+1}] = r$, for all t. In solving the above difference equation forward, that is, in replacing p_{t+1} with $E_{t+1}[p_{t+2} + d_{t+2}]/(1 + r)$ in Eq. (1) versus Eq. (2) below and then p_{t+2} and so on, and using the law of iterated expectations, one obtains after $T - t - 1$ iterations

$$p_t = E_t\left[\sum_{\tau=1}^{T-1} \frac{1}{(1+r)^\tau} d_{t+\tau}\right] + E_t\left[\frac{1}{(1+r)^{T-t}} p_T\right] .$$

The equilibrium price is given by the expected discounted value of the future dividend stream paid from $t + 1$ to T plus the expected discounted value of the price at T. For *securities with finite maturity*, the price after maturity, say T, is zero, $p_T = 0$. Hence, the price of the asset, p_t, is unique and simply coincides with the expected future discounted dividend stream until maturity. Put differently, finite horizon bubbles cannot arise as long as rational investors are unconstrained from selling the desired number of shares in all future contingencies. For *securities with infinite maturity*, $T \to \infty$, the price p_t only coincides with the expected discounted value of the future dividend stream, call it fundamental value, v_t, if the so-called transversality condition, $\lim_{T\to\infty} E_t\left[\frac{1}{(1+r)^{T-t}} p_T\right] = 0$, holds. Without imposing the transversality condition, $p_t = v_t$ is only one of many possible prices that solve the above expectational difference equation. Any price $p_t = v_t + b_t$, decomposed in the fundamental value, v_t, and a bubble component, b_t, such that

$$b_t = E_t\left[\frac{1}{(1+r)} b_{t+1}\right], \tag{2}$$

is also a solution. Equation (2) versus Eq. (1) needs to be made consistent. Equation (2) highlights that the bubble component b_t has to 'grow' in expectations exactly at a rate of r. A nice example of these 'rational bubbles' is provided in Blanchard and Watson (1982), where the bubble

persists in each period only with probability π and bursts with probability $(1 - \pi)$. If the bubble continues, it has to grow in expectation by a factor $(1 + r)/\pi$. This faster bubble growth rate (conditional on not bursting) is necessary to achieve an expected growth rate of r. In general, the bubble component may be stochastic. A specific example of a stochastic bubble is an intrinsic bubble, where the bubble component is assumed to be deterministically related to a stochastic dividend process.

The fact that any bubble has to grow at an expected rate of r allows one to eliminate many potential rational bubbles. For example, a positive bubble cannot emerge if there is an upper limit on the size of the bubble. That is, for example, the case with potential bubbles on commodities with close substitutes. An ever-growing 'commodity bubble' would make the commodity so expensive that it would be substituted with some other good. Similarly, a bubble on a non-zero net supply asset cannot arise if the required return r exceeds the growth rate of the economy, since the bubble would outgrow the aggregate wealth in the economy. Hence, bubbles can only exist in a world in which the required return is lower than or equal to the growth rate of the economy. In addition, rational bubbles can persist if the pure existence of the bubble enables trading opportunities that lead to a different equilibrium allocation. Fiat money in an overlapping generations (OLG) model is probably the most famous example of such a bubble. The intrinsic value of fiat money is zero, yet it has a positive price. Moreover, only when the price is positive, does it allow wealth transfers across generations (that might not even be born yet). A negative bubble, $b_t < 0$, on a limited-liability asset cannot arise since the bubble would imply that the asset price has to become negative in expectation at some point in time. This result, together with Eq. (2), implies that if the bubble vanishes at any point it has to remain zero from that point onwards. That is, rational bubbles can never emerge within an asset-pricing model; they must already be present when the asset starts trading.

Empirically testing for rational bubbles under symmetric information is a challenging task. The literature has developed three types of tests: regression analysis, variance bounds tests and experimental tests. Initial tests proposed by Flood and Garber (1980) exploit the fact that bubbles cannot start within a rational asset-pricing model and hence at any point in time the price must have a non-zero part that grows at an expected rate of r. However using this approach, inference is difficult due to an exploding regressor problem. That is, as time t increases, the regressor explodes and the coefficient estimate relies primarily on the most recent data points. More precisely, the ratio of the information content of the most recent data point to the information content of all previous observations never goes to zero. This implies that as time t increases, the time series sample remains essentially small and the central limit theorem does not apply. Diba and Grossman (1988) test for bubbles by checking whether the stock price is more explosive than the dividend process. Note that if the dividend process follows a linear unit-root process (for example, a random walk), then the price process has a unit root as well. However the change in price, Δp_t, and the spread between the price and the discounted expected dividend stream, $p_t - d_t/r$, are stationary under the no-bubbles hypothesis. That is, p_t and d_t/r are co-integrated. Diba and Grossman test this hypothesis using a series of unit root tests, autocorrelation patterns, and co-integration tests. They conclude that the no-bubble hypothesis cannot be rejected. However, Evans (1991) shows that these standard linear econometric methods may fail to detect the explosive nonlinear patterns of periodically collapsing bubbles. West (1987) proposes a different test that exploits the fact that one can estimate the parameters needed to calculate the expected discounted value of dividends in two different ways. One way of estimating them is not affected by the bubble, the other is. Note that the accounting identity (1) can be rewritten as $p_t = \frac{1}{1+r}\left(p_{t+1} + d_{t+1}\right) - \frac{1}{1+r}\left(p_{t+1} + d_{t+1} - E_t\left[p_{t+1} + d_{t+1}\right]\right)$. Hence, in an instrumental variables regression of p_t on $p_{t+1} + d_{t+1}$ – using for example d_t as an instrument – one obtains an estimate for r that is independent of the existence of a rational bubble.

Second, if, for example, the dividend process follows a stationary AR(1) process, $d_{t+1} = \varphi d_t + \eta_{t+1}$, with independent noise η_{t+1}, one can easily estimate φ Furthermore, the expected discounted value of future dividends is $v_t = (\varphi/(1 + r - \varphi))d_t$. Hence, under the null-hypothesis of no bubble, that is $p_t = v_t$, the coefficient estimate of the regression of p_t on d_t provides a second estimate of $\varphi/(1 + r - \varphi)$ In a final step, West uses a Hausman specification test to test whether both estimates coincide. He finds that the US stock market data usually reject the null hypothesis of no bubble.

Excessive volatility in the stock market seems to provide further evidence in favour of stock market bubbles. LeRoy and Porter (1981) and Shiller (1981) introduced variance bounds that indicate that the stock market is too volatile to be justified by the volatility of the discounted dividend stream. However, the variance bounds test is controversial (see, for example, Kleidon 1986). Also, this test, as well as all the aforementioned bubble tests, assumes that the required expected returns, r, are constant over time. In a setting in which the required expected returns can be time-varying, the empirical evidence favouring excess volatility is less clear-cut. Furthermore, time-varying expected returns can also rationalize the long-horizon predictability of stock returns. For example, a high price–dividend ratio predicts low subsequent stock returns with a high R^2 (Campbell and Shiller 1988).

Finally, it is important to recall that the theoretical arguments that rule out rational bubbles as well as several empirical bubble tests rely heavily on backward induction. Since a bubble cannot grow from time T onwards, there cannot be a bubble of this size at time $T - 1$, which rules out this bubble at $T - 2$, and so on. However, there is ample *experimental evidence* that individuals violate the backward induction principle. Most convincing are experiments on the centipede game (Rosenthal 1981). In this simple game, two players alternatively decide whether to continue or stop the game for a finite number of periods. On any move, a player is better off stopping the game than continuing if the other player stops immediately afterwards, but is worse off stopping than continuing if the other player continues afterwards. This game has only a single subgame perfect equilibrium that follows directly from backward induction reasoning. Each player's strategy is to stop the game whenever it is his or her turn to move. Hence, the first player should immediately stop the game and the game should never get off the ground. However, in experiments players initially continue to play the game – a violation of the backward induction principle (see for example, McKelvey and Palfrey 1992). These experimental findings question the theoretical reasoning used to rule out rational bubbles under symmetric information. More experimental evidence on bubbles in general is provided in the final section.

In a rational bubble setting an investor only holds a bubble asset if the bubble grows in expectations ad infinitum. In contrast, in the following models an investor might hold an overpriced asset if he thinks he can resell it in the future to a less informed trader or someone who holds biased beliefs. In Kindleberger's (2000) terms, the investor thinks he can sell the asset to a greater fool.

Asymmetric Information Bubbles

Asymmetric information bubbles can occur in a setting in which investors have different information, but still share a common prior distribution. In these models prices have a dual role: they are an index of scarcity and informative signals, since they aggregate and partially reveal other traders' aggregate information (see for example Brunnermeier 2001 for an overview). In contrast to the symmetric information case, the presence of a bubble need not be commonly known. For example, it might be the case that everybody knows the price exceeds the value of any possible dividend stream, but it is not the case that everybody knows that all the other investors also know this fact. It is this lack of higher-order mutual knowledge that makes it possible for finite bubbles to exist under certain necessary conditions (Allen et al. 1993). First, it is crucial that investors remain asymmetrically informed even after inferring information from prices and net trades. This

implies that prices cannot be fully revealing. Second, investors must be constrained from (short) selling their desired number of shares in at least one future contingency for finite bubbles to persist. Third, it cannot be common knowledge that the initial allocation is interim Pareto efficient, since then it would be commonly known that there are no gains from trade and hence the buyer of an overpriced 'bubble asset' would be aware that the rational seller gains at his expense (Tirole 1982). In other words, there have to be gains from trade or at least some investors have to think that there might be gains from trade. There are various mechanisms that lead to these. For example, fund managers who invest on behalf of their clients can gain from buying overpriced bubble assets, since trading allows them to fool their clients into believing that they have superior trading information. A fund manager who does not trade would reveal that he does not have private information. Consequently, bad fund managers churn bubbles at the expense of their uninformed client investors (Allen and Gorton 1993). Furthermore, fund managers with limited liability might trade bubble assets due to classic risk-shifting incentives, since they participate on the potential upside of a trade but not on the downside risk.

Bubbles Due to Limited Arbitrage

Bubbles due to limited arbitrage arise in models in which rational, well-informed and sophisticated investors interact with behavioural market participants whose trading motives are influenced by psychological biases. Proponents of the 'efficient markets hypothesis' argue that bubbles cannot persist since well-informed sophisticated investors will undo the price impact of behavioural non-rational traders. Thus, rational investors should go against the bubble even before it emerges. The literature on limits to arbitrage challenges this view. It argues that bubbles can persist, and provides three channels that prevent rational arbitrageurs from fully correcting the mispricing. First, *fundamental risk* makes it risky to short a bubble asset since a subsequent positive shift in fundamentals might *ex post* undo the initial overpricing. Risk aversion limits the aggressiveness of rational traders if close substitutes and close hedges are unavailable. Second, rational traders also face *noise trader risk* (DeLong et al. 1990). Leaning against the bubble is risky even without fundamental risk, since irrational noise traders might push up the price even further in the future and temporarily widen the mispricing. Rational traders with short horizons care about prices in the near future in addition to the long-run fundamental value and only partially correct the mispricing. For example, in a world with delegated portfolio management, fund managers are often concerned about short-run price movements, because temporary losses instigate fund outflows (Shleifer and Vishny 1997). A temporary widening of the mispricing and the subsequent outflow of funds force fund managers to unwind their positions exactly when the mispricing is the largest. Anticipating this possible scenario, mutual fund managers trade less aggressively against the mispricing. Similarly, hedge funds face a high flow-performance sensitivity, despite some arrangements designed to prevent outflows (for example, lock-up provisions). Third, rational traders face *synchronization risk* (Abreu and Brunnermeier 2002, 2003). Since a single trader alone cannot typically bring the market down by himself, coordination among rational traders is required and a synchronization problem arises. Each rational trader faces the following trade-off: if he attacks the bubble too early, he forgoes profits from the subsequent run-up caused by behavioural momentum traders; if he attacks too late and remains invested in the bubble asset, he will suffer from the subsequent crash. Each trader tries to forecast when other rational traders will go against the bubble. Timing other traders' moves is difficult because traders become sequentially aware of the bubble, and they do not know where in the queue they are. Because of this 'sequential awareness', it is never common knowledge that a bubble has emerged. It is precisely this lack of common knowledge that removes the bite of the standard backward induction argument. Since there is no commonly known point in time from which one could start backward induction, even finite horizon bubbles can persist.

The other important message of the theoretical work on synchronization risk is that relatively insignificant news events can trigger large price movements, because even unimportant news events allow traders to synchronize their sell strategies. Unlike the earlier limits to arbitrage models, in which rational traders do not trade aggressively enough to completely eradicate the bubble but still short an overpriced bubble asset, in Abreu and Brunnermeier (2003) rational traders prefer to ride the bubble rather than attack it. The incentive to ride the bubble stems from a predictable 'sentiment' in the form of continuing bubble growth.

Empirically, there is supportive evidence in favour of the 'bubble-riding hypothesis'. For example, between 1998 and 2000 hedge funds were heavily tilted towards highly priced technology stocks (Brunnermeier and Nagel 2004). Contrary to the efficient markets hypothesis, hedge funds were not a price-correcting force even though they are among the most sophisticated investors and are arguably closer to the ideal of 'rational arbitrageurs' than any other class of investors. Similarly, Temin and Voth (2004) document that Hoares Bank was profitably riding the South Sea bubble in 1719–20, despite giving numerous indications that it believed the stock to be overvalued. Many other investors, including Isaac Newton, also tried to ride the South Sea bubble but with less success. Frustrated with his trading experience, Isaac Newton concluded 'I can calculate the motions of the heavenly bodies, but not the madness of people' (Kindleberger 2005, p. 41).

Heterogeneous Beliefs Bubbles

Bubbles can also emerge when investors have heterogeneous beliefs and face short- sale constraints. Investors' beliefs are heterogeneous if they start with different prior belief distributions that can be due to psychological biases. For example, if investors are overconfident about their own signals, they have a different prior distribution (with lower variance) about the signals' noise term. Investors with non-common priors can agree to disagree even after they share all their information. Also, in contrast to an asymmetric information setting, investors do not try to infer other traders' information from prices. Combining heterogeneous beliefs with short-sale constraints can result in overpricing since optimists push up the asset price, while pessimists cannot counterbalance it since they face short-sale constraints (Miller 1977). Ofek and Richardson (2003) link this argument to the Internet bubble of the late 1990s. In a dynamic model, the asset price can even exceed the valuation of the most optimistic investor in the economy. This is possible, since the currently optimistic investors – the current owners of the asset – have the option to resell the asset in the future at a high price whenever they become less optimistic. At that point other traders will be more optimistic, and hence be willing to buy the asset since optimism is assumed to oscillate across different investor groups (Harrison and Kreps 1978). It is essential that less optimistic investors, who would like to short the asset, are prevented from doing so by the short-sale constraint. Heterogeneous belief bubbles are accompanied by large trading volume and high price volatility (Scheinkman and Xiong 2003).

Experimental Evidence

Many theoretical arguments in favour of or against bubbles are difficult to test with (confounded) field data. Laboratory experiments have the advantage that they allow the researcher to isolate and test specific mechanisms and theoretical arguments. For example, the aforementioned experimental evidence on centipede games questions the validity of backward induction. There is a large and growing literature that examines bubbles in a laboratory setting. For example, Smith et al. (1988) study a double-auction setting, in which a risky asset pays a uniformly distributed random dividend of $d \in \{0, d_1, d_2, d_3\}$ in each of the 15 periods. Hence, the fundamental value for a risk-neutral trader is initially $15\sum_i \frac{1}{4} d_i$ and declines by $\sum_i \frac{1}{4} d_i$ in each period. Even though there is no asymmetric information and the probability distribution is

commonly known, there is vigorous trading, and prices initially rise despite the fact that the fundamental value steadily declines. More specifically, the time-series of asset prices in the experiments are characterized by three phases. An initial boom phase is followed by a period during which the price exceeds the fundamental value, before the price collapses towards the end. These findings are in sharp contrast to any theoretical prediction and seem very robust across various treatments. A string of subsequent articles show that bubbles still emerge after allowing for short sales, after introducing trading fees, and when using professional business people as subjects. Only the introduction of futures markets and the repeated experience of a bubble reduce the size of the bubble.Researchers have speculated that bubbles emerge because each trader hopes to outwit others and to pass the asset on to some less rational trader in the final trading rounds. However, more recent research has revealed that the lack of common knowledge of rationality is not the cause of bubbles. Even when investors have no resale option and are forced to hold the asset until the end, bubbles still emerge (Lei et al. 2001).

In summary, the literature on bubbles has taken giant strides since the 1970s that led to several classes of models with distinct empirical tests. However, many questions remain unresolved. For example, we do not have many convincing models that explain when and why bubbles start. Also, in most models bubbles burst, while in reality bubbles seem to deflate over several weeks or even months. While we have a much better idea of why rational traders are unable to eradicate the mispricing introduced by behavioural traders, our understanding of behavioural biases and belief distortions is less advanced. From a policy perspective, it is interesting to answer the question whether central banks actively try to burst bubbles. I suspect that future research will place greater emphasis on these open issues.

See Also

- ▶ Behavioural Finance
- ▶ Kindleberger, Charles P. (1910–2003)
- ▶ South Sea Bubble
- ▶ Speculative Bubbles
- ▶ Tulipmania

Bibliography

Abreu, D., and M.K. Brunnermeier. 2002. Synchronization risk and delayed arbitrage. *Journal of Financial Economics* 66: 341–360.

Abreu, D., and M.K. Brunnermeier. 2003. Bubbles and crashes. *Econometrica* 71: 173–204.

Allen, F., and G. Gorton. 1993. Churning bubbles. *Review of Economic Studies* 60: 813–836.

Allen, F., S. Morris, and A. Postlewaite. 1993. Finite bubbles with short sale constraints and asymmetric information. *Journal of Economic Theory* 61: 206–229.

Blanchard, O.J., and M.W. Watson. 1982. Bubbles, rational expectations, and financial markets. In *Crisis in the economic and financial structure*, ed. P. Wachtel. Lexington: Lexington.

Brunnermeier, M.K. 2001. *Asset pricing under asymmetric information: Bubbles, crashes, technical analysis and herding*. Oxford: Oxford University Press.

Brunnermeier, M.K., and S. Nagel. 2004. Hedge funds and the technology bubble. *Journal of Finance* 59: 2013–2040.

Campbell, J.Y., and R.J. Shiller. 1988. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1: 195–228.

DeLong, J.B., A. Shleifer, L.H. Summers, and R.J. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98: 703–738.

Diba, B.T., and H.I. Grossman. 1988. The theory of rational bubbles in stock prices. *Economic Journal* 98: 746–754.

Evans, G.W. 1991. Pitfalls in testing for explosive bubbles in asset prices. *American Economic Review* 81: 922–930.

Flood, R.P., and P.M. Garber. 1980. Market fundamentals versus price-level bubbles: The first tests. *Journal of Political Economy* 88: 745–770.

Harrison, J.M., and D. Kreps. 1978. Speculative investor behavior in a stock market with heterogeneous expectations. *Quarterly Journal of Economics* 89: 323–336.

Kindleberger, C.P. 2005. *Manias, panics and crashes: A history of financial crises*. 5th ed. New york: Wiley.

Kleidon, A.W. 1986. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94: 953–1001.

Lei, V., C.N. Noussair, and C.R. Plott. 2001. Nonspeculative bubbles in experimental asset markets: Lack of common knowledge of rationality vs. actual irrationality. *Econometrica* 69: 831–859.

LeRoy, S.F., and R.D. Porter. 1981. The present value relation: Tests based on implied variance bounds. *Econometrica* 64: 555–574.

McKelvey, R.D., and T.R. Palfrey. 1992. An experimental study of the centipede game. *Econometrica* 60: 803–836.
Miller, E.M. 1977. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32: 1151–1168.
Ofek, E., and M. Richardson. 2003. *DotCom mania: The rise and fall of Internet stocks*, Working paper no. FIN-01-037 58(3), 1113–1138. New York University, Stern School.
Rosenthal, R. 1981. Games of perfect information, predatory pricing, and the chain- store paradox. *Journal of Economic Theory* 25: 92–100.
Scheinkman, J., and W. Xiong. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 111: 1183–1219.
Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.
Shleifer, A., and R.W. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 52: 35–55.
Smith, V.L., G.L. Suchanek, and A.W. Williams. 1988. Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica* 56: 1119–1151.
Temin, P., and H.-J. Voth. 2004. Riding the south sea bubble. *American Economic Review* 94: 1654–1668.
Tirole, J. 1982. On the possibility of speculation under rational expectations. *Econometrica* 50: 1163–1182.
West, K.D. 1987. A specification test for speculative bubbles. *Quarterly Journal of Economics* 102: 553–580.

Bubbles in History

Charles P. Kindleberger
Organization Name, City, UK

Keywords

Bank lending; Booms; Bubbles; German inflation (1923); Hedge finance; Kindleberger, C. P.; Law, J.; Mississippi bubble; Ponzi finance; Railway mania; Rational expectations; South Sea bubble; Speculative finance; Tulipmania

JEL Classifications

G1

A bubble may be defined loosely as a sharp rise in price of an asset or a range of assets in a continuous process, with the initial rise generating expectations of further rises and attracting new buyers – generally speculators interested in profits from trading in the asset rather than its use or earning capacity. The rise is usually followed by a reversal of expectations and a sharp decline in price often resulting in financial crisis. A boom is a more extended and gentler rise in prices, production, and profits than a bubble and may be followed by crisis, sometimes taking the form of a crash (or panic) or alternatively by a gentle subsidence of the boom without crisis.

Bubbles have existed historically, at least in the eyes of contemporary observers, as well as booms so intense and excited that they have been called 'manias'. The most notable bubbles were the Mississippi bubble in Paris in 1719–1720, set in motion by John Law, founder of the *Banque Générale* and the *Banque Royale*, and the contemporaneous and related South Sea bubble in London. Most famous of the manias were the Tulip mania in Holland in 1636 and the Railway mania in England in 1846–1847. It is sometimes debated whether a particular sharp rise and fall in prices, such as the German hyperinflation from 1920 to 1923, the rise and fall in commodity and share prices in London and New York in 1919–1921, and the rise of gold of $850 an ounce in 1982 and its subsequent fall to the $350 level were or were not bubbles. Some theorists go further and question whether bubbles are possible with rational markets, which they assume exist (see, e.g., Flood and Garber 1980).

Rational expectations theory holds that prices are formed within the limits of available information by market participants using standard economic models appropriate to the circumstances. As such, it is claimed, market prices cannot diverge from fundamental values unless the information proves to have been widely wrong. The theoretical literature uses the assumption of the market having one mind and one purpose, whereas it is observed historically that market participants are often moved by different purposes, operate with different wealth and information, and calculate within different time horizons. In early railway investment, for example, initial investors were persons doing business along the

rights of way who sought benefits from the railroad for their other concerns. They were followed by a second group of investors interested in the profits the railroad would earn, and by a third group, made up of speculators who, seeing the rise in the railroad's shares, borrowed money or paid for the initial instalments with no intention of completing the purchase, to make a profit on resale.

The objects of speculation resulting in bubbles or booms and ending in numerous cases, but not all, in financial crisis, change from time to time and include commodities, domestic bonds, domestic shares, foreign bonds, foreign shares, urban and suburban real estate, rural land, leisure homes, shopping centres, Real Estate Investment Trusts, 747 aircraft, supertankers, so-called collectibles such as paintings, jewellery, stamps, coins, antiques, etc. and, most recently, syndicated bank loans to developing countries. Within these relatively broad categories, speculation may fix on particular objects – insurance shares, South American mining stocks, cotton-growing land, Paris real estate, Post-Impressionist art, and the like.

At the time of writing, the theoretical literature has yet to converge on an agreed definition of bubbles and on whether they are possible. Virtually the same authors who could not reject the no-bubbles hypothesis in the German inflation of 1923 one year, managed to do so a year later (Flood and Garber 1980). Another pair of theorists has demonstrated mathematically that rational bubbles can exist after putting aside "irrational bubbles" on the grounds not of their nonexistence but of the difficulty of the mathematics involved (Blanchard and Watson 1982).

Short of bubbles, manias, and irrationality are periods of euphoria which produce positive feedback, price increases greater than justified by market fundamentals, and booms of such dimensions as to threaten financial crisis, with possibilities of a crash or panic. Minsky (1982a, b) has discussed how after an exogenous change in economic circumstances has altered profit opportunities and expectations, bank lending can become increasingly lax by rigorous standards. Critical exception has been taken to his taxonomy dividing bank lending into hedge finance, to be repaid out of anticipated cash flows; speculative finance, requiring later refinancing because the term of the loan is less than the project's payoff; and Ponzi finance, in which the borrower expects to pay off his loan with the proceeds of sale of an asset. It is objected especially that Carlo Ponzi was a swindler and that many loans of the third type, for example, those to finance construction, are entirely legitimate (Flemming et al. 1982). Nonetheless, the suggestion that lending standards grow more lax during a boom and that the banking system on that account becomes more fragile has strong historical support. It is attested, and the contrary rational-expectations view of financial markets is falsified, by the experience of such a money and capital market as London having successive booms, followed by crisis, the latter in 1810, 1819, 1825, 1836, 1847, 1857, 1866, 1890, 1900, and 1921 – a powerful record of failing to learn from experience (Kindleberger 1978).

See Also

► Tulipmania

Bibliography

Blanchard, O., and M.W. Watson. 1982. Bubbles, rational expectations and financial markets. In *Crises in the economic and financial structure*, ed. P. Wachtel. Lexington: Heath.

Flemming, J.S., R.W. Goldsmith, and J. Melitz. 1982. Comment. In *Financial crises: Theory, history and policy*, ed. C.P. Kindleberger and J.-P. Laffargue. Cambridge: Cambridge University Press.

Flood, R.P., and P.M. Garber. 1980. Market fundamentals versus price-level bubbles: The first tests. *Journal of Political Economy* 88: 745–770.

Kindleberger, C.P. 1978. *Manias, panics and crashes: A history of financial crises*. New York: Basic Books.

Minsky, H.P. 1982a. *Can 'it' happen again?: Essays on instability and finance*. Armonk: Sharpe.

Minsky, H.P. 1982b. The financial instability hypothesis. In *Financial crises: Theory, history and policy*, ed. C.P. Kindleberger and J.-P. Laffargue. Cambridge: Cambridge University Press.

Buchanan, David (1779–1848)

Andrew Skinner

Buchanan was born in Montrose, the eldest son of David Buchanan, the renowned printer, publisher and amateur literary scholar. Unlike his father, David the younger did not attend university, but entered the family business. Primarily interested in economics, geography and statistics, Buchanan is generally regarded as a journalist and writer, but also as a 'Scottish economist' (*Encyclopedia of the Social Sciences,* 1935, iii. 27). Buchanan's career amply justifies all of these claims.

Invited by Francis Horner and Francis Jeffrey to act as editor for the short-lived *Weekly Register* in 1808, Buchanan moved to the *Caledonian Mercury* two years later and remained in this post until 1827. In the same year he became editor of the *Edinburgh Courant,* a position he held until his sudden death in 1848.

In 1835 Buchanan helped to compile the *Edinburgh Geographical Atlas,* and made a number of contributions to the *Edinburgh Gazetteer.* He also contributed pieces on geography and statistics to the seventh edition of the *Encyclopaedia Britannica* (1842), which were acknowledged in the preface. But the bulk of Buchanan's output was on economics, with numerous articles appearing in Cobbett's *Political Register* and in the *Edinburgh Review.* The latter in particular carried pieces on 'Lord Henry Petty's plan of Finance' (1807), 'Wheatley on money and finance' (1807), 'Spence on agriculture and commerce' (with Francis Jeffrey, 1809), the Corn Laws (1815), and 'Corn and money' (1816).

This growing interest in economic subjects prepared Buchanan for his critical, annotated edition of the *Wealth of Nations* (1814), which in turn paved the way for his *Observations on the Subjects Treated of in Dr. Smith's Inquiry* published in the same year. In the Introduction to the latter work Buchanan set Smith's achievement in the context of the work done by Sir James Steuart and the physiocrats. While expressing qualified admiration for both, Buchanan noted that the *Wealth of Nations* 'is a great display of reason on the business of the world; touching society in all its essential relations, containing lessons for government as well as for common life, and embracing subjects formerly placed without the limits of philosophy' (1814, p. viii).

Yet Dr Smith had 'not published a perfect work'. The critical 'dissertations' which follow supplement the notes with the intention of correcting 'what is amiss' (p. xv).

Less successful in his treatment of Ricardo, Buchanan elaborated on the determinants of price and criticized Smith's theory of rent. Other subjects covered included metallic money and paper currency, wages, stock, productive and unproductive labour, the progress of opulence, the Corn Laws, commercial treaties, defence, public debt and the East India Company.

Buchanan included a section on taxation and went on to publish an *Inquiry into the Taxation and Commercial Policy of Great Britain* (1844) which subsequently attracted some critical acclaim.

In 1852 Buchanan was described as a man of 'unobtrusive habits, mild and gentle in his demeanour, and held in high respect by all who had an opportunity of forming an estimate of his character' (Anderson 1863, p. 481).

Selected Works

1814. (ed.) *Inquiry into the nature and causes of the wealth of nations,* in three volumes; to which is added *Observations on the Subjects Treated of in Dr. Smith's Inquiry* (1814), Edinburgh/London: Oliphant, Waugh & Innes/John Murray.

1844. *Inquiry into the taxation and commercial policy of Great Britain, with observation on the principles of currency and of exchangeable value.* Edinburgh: W. Tait.

References

Anderson, W. 1863. *The Scottish nation*. Edinburgh: A. Fullarton.
Buchanan, David. 1886. *Dictionary of national biography*, vol. VII. London: Smith, Elder & Co.
Buchanan, David. 1935. *Encyclopedia of the social sciences*, vol. III. New York: Macmillan.
Buchanan, David. 1894. In *Dictionary of political economy*, vol. I, ed. R.H. Inglis Palgrave. London: Macmillan.
Houghton, W.E. (ed.). 1966. *Wellesley index to Victorian Periodicals, 1824–1900*. Toronto: Toronto University Press.

Buchanan, James M. (Born 1919)

Richard E. Wagner

Abstract

The customary Anglo-Saxon approach to public finance treats the state as exogenous to the economic process, which restricts public finance to the study of market-based reactions to exogenous fiscal impositions. In contrast, Buchanan has cultivated an approach to public finance that incorporates the state into the economic process. The domain of fiscal analysis is thus expanded in two directions. One direction, public choice, involves the study of the effect of political institutions on collective choices. The other direction, constitutional political economy, involves the emergence of and changes in political institutions.

Keywords

Bequest motive; Buchanan, J.; Constitutional political economy; Constitutional vs. post-constitutional behaviour; Decision costs; External costs; Fiscal institututions; Free rider problem; Knight, F.; Majoritarianism; Progressive and regressive taxation; Public choice; Public debt; Ricardian equivalence; State, size of; Strategic behaviour; Unanimity; Voting rules; Wicksell, K.

JEL Classifications

B31

James M. Buchanan was awarded the 1986 Nobel Memorial Prize in Economic Science for his seminal role in developing 'the contractual and constitutional bases for the theory of economic and political decision-making'.

Buchanan spent his boyhood in rural Tennessee near Murfreesboro. After receiving Bachelor's and Master's degrees from Middle Tennessee State College and the University of Tennessee respectively, he entered the US Navy in 1941. After completing his naval service in the Pacific, Buchanan enrolled at the University of Chicago in 1946, receiving his Ph.D. in 1948. He has spent the preponderance of his academic career at three Virginia universities: the University of Virginia (1956–68), Virginia Polytechnic Institute (1969–83), and George Mason University (since 1983). Buchanan has been a truly prolific scholar throughout this period, as shown by the 20 volumes of his collected works published by Liberty Fund; moreover, he has continued his scholarly work at full speed since the completion of that collection in 2001.

The Nobel citation referred to above identifies two predominant strains within Buchanan's scholarly oeuvre. One of these is the theory of public choice, which entails the application of economic theorizing to politics. The other is constitutional political economy, which explores the relationship between constitutional rules and political outcomes. While Buchanan's body of work also contains numerous contributions to economic theory and methodology, which by themselves would have constituted a significant scholarly career, this short article focuses exclusively on Buchanan's approach to public choice and constitutional political economy.

Precursory Influences

While Buchanan has been creative as well as prolific, he has nonetheless been inspired by, and has built upon, the contributions of others.

Buchanan has acknowledged these precursory influences numerous times, particularly in his autobiographical *Better than Plowing,* where he identifies three sources of primary influence on his work.

The primary precursors to Buchanan's public choice theorizing were a set of Italian scholars, among them Antonio De Viti De Marco, Maffeo Pantaleoni, and Luigi Einaudi, who developed a unique orientation towards public finance between the 1880s and the 1930s. Where Anglo-Saxon scholars treated the state as outside the economy, the Italians sought to incorporate political outcomes into the economic process. For instance, much Anglo-Saxon fiscal scholarship sought to develop norms regarding the desirable degree of tax progressivity, as illustrated by various sacrifice theories of taxation. By contrast, the Italians sought to explain the actual structure of taxation independently of normative concern, and to do so with reference to the same categories of utility and cost as they invoked to explain market outcomes. This Italian orientation of sober realism towards political processes was central to the later development of public choice theorizing. For instance, in his foreword to the German translation of Amilcare Puviani's 1903 treatise on fiscal illusion, *Teoria della illusione finanziaria,* Gunter Schmölders observed that 'over the last century Italian public finance has had an essentially political science character. . . . This work [Puviani] is a typical product of Italian public finance. . . . Above all, it is the science of public finance combined with fiscal politics, in many places giving a good fit with reality' (Puviani 1960). The Italians were thoroughgoing realists and not romantic idealists, and it was a short distance from their initial formulations to what subsequently became known as public choice.

The sober realism of the Italians implied, in keeping with the general equilibrium theorizing of the time, that actual fiscal outcomes were to be explained as equilibrium outcomes. If so, it might seem as though fiscal theorizing offered no coherent vantage point from which to pursue any programme of fiscal reform. Yet Buchanan has always sought to use fiscal knowledge as an instrument of fiscal reform. It was Knut Wicksell who provided Buchanan with the vehicle for combining his sober realism with his interest in reform. Buchanan's constitutional emphasis can be traced to the second of Wicksell's three essays in *Finanztheoretische Untersuchungen,* which Buchanan translated as 'A New Theory of Just Taxation', in *Classics in the Theory of Public Finance,* edited by Richard Musgrave and Alan Peacock. From Wicksell, Buchanan derived two themes that informed his work thereafter. One theme was the treatment of unanimous consent and not majority approval as the normative benchmark for appraising political outcomes. The other theme was a distinction between constitutional politics, where institutional rules are selected, and post-constitutional politics, where particular outcomes emerge. Wicksell's treatment of two distinct levels of political activity led to Buchanan's articulation of a constitutional political economy, wherein political reform was a matter of changing the rules that govern the game, as distinct from changing the strategies of play within a game.

While Wicksell and the Italians cover the two themes mentioned in Buchanan's Nobel citation, any mention of precursory influences would be remiss without including Frank Knight, whom Buchanan initially encountered during his student days at the University of Chicago. Knight's influence on Buchanan is not so much one of particular ideas as of general attitude and orientation towards a scholar's life and work. From Knight, Buchanan carried forward the belief that no doctrine or authority should be treated as sacrosanct and above challenge. Everyone else may say that something is true, but this doesn't mean they are right; there may be many pretentious emperors walking around naked. Buchanan's work has also demonstrated the same multidisciplinary character that was prominent in Knight's work. For Buchanan, as for Knight, economic theorizing was not self-contained, but had points of contact throughout the humane studies, which led to a style of theorizing wherein Buchanan, like Knight, continually makes contact with such related fields of inquiry as law, ethics, history, philosophy, and politics.

From Italian Public Finance to Public Choice

The Italian approach to public finance treated the state as an entity whose actions conformed to the same principles of marginal utility as the actions of other economic participants. The Italians did not seek to advance statements concerning how large the state should be in order to promote some vision of social welfare. They sought instead to offer coherent explanations about the actual size of the state. At the level of formal analysis, this meant that the state would expand until the marginal utility from state-provided services equalled the marginal utility from market-supplied services. To be sure, the Italians recognized the numerous problems of aggregation that were involved in making such statements. In response, they developed a variety of models regarding just whose utility was driving the equilibrium. Where some models treated the state as a cooperative enterprise that worked to the benefit of all, others treated the state as an entity that promoted the advantage of ruling classes. In any case, it was a small step from the Italian fiscal theorizing to the public choice theorizing that began to take shape in the 1960s, as elaborated in Richard Wagner (2003).

Perhaps the best place to see the Italian influence on public choice is Buchanan's 1967 treatise *Public Finance in Democratic Process,* which was written at a time when 'public choice' was not yet a term of scholarly identification. Buchanan starts that book by noting the narrow and limited scope of Anglo-Saxon approaches to public finance, wherein public finance is concerned only with explaining market-based reactions to exogenously imposed taxes and expenditures. On the tax side of the budget, for instance, a progressive income tax with several brackets of rising marginal rates might be replaced by a degressive tax where a single marginal rate is imposed above some initial exemption. The task of the fiscal scholar would be to explain the impact of such an exogenous tax shift on such things as the amount of labour people supply, the amount of underground economic activity they undertake, and the amount of taxable income they earn. Alternatively, on the expenditure side of the budget, an appropriation might be made to finance a highway. The task of fiscal analysis would be to analyse the market-based reactions to the highway. For instance, land rents near highway exits might rise due to the reduction in travel time that resulted. Whatever the particular topic examined, the analytical task of Anglo-Saxon public finance has everything to do with explaining market-based reactions to exogenously imposed fiscal measures and has nothing to do with explaining state budgets and fiscal institutions.

In treating state budgets as exogenous to fiscal inquiry, the Anglo-Saxon orientation towards public finance ignored two large areas of possible inquiry, both of which Buchanan explores in *Public Finance in Democratic Process.* One ignored area is the ability of fiscal institutions to influence budgetary outcomes and not just market outcomes. This topic occupies the first part of *Public Finance in Democratic Process*, and the analyses presented there were early illustrations of public choice theorizing. The second ignored area is the choice or emergence of fiscal institutions. This topic occupies the second part of *Public Finance in Democratic Process*, and the analyses presented there were harbingers of subsequent work in constitutional political economy.

Buchanan gives several illustrations in *Public Finance in Democratic Process* of how fiscal institutions and arrangements might influence fiscal outcomes, of which I mention three. First, Buchanan examines the possible budgetary consequences of a choice between general-fund financing and tax earmarking. Under the former practice, tax revenues accrue to a general fund from which various appropriations are made; under the latter practice, specific taxes are earmarked to finance particular services. Buchanan suggests that general-fund financing is a form of tie-in sale that might bring about a budgetary shift in favour of services in relatively elastic demand.

Second, Buchanan examines the possible budgetary consequences of the withholding of income taxes. His analysis in this case is related to claims about fiscal illusion or perception. Buchanan

argues that individual perceptions about the costliness of public output depend on the manner in which tax extractions are made. Perhaps the most open and direct manner of paying for public output would be for people to write monthly checks to government, just as they pay their utility bills. Buchanan explores the possibility that withholding may create some tendency for individuals to perceive the cost of government to be less than it would otherwise be, which should in turn lead to some increase in the size of government.

Third, Buchanan examines the effect of public debt on budgetary outcomes, a topic that he initially explored in *Public Principles of Public Debt* and to which he returned in *Democracy in Deficit* (co-authored with Richard Wagner). The principle of Ricardian equivalence holds that tax finance and debt finance are identical. In the aggregate, this is true as a simple matter of double-entry accounting. If $1 million of tax revenue is replaced by public borrowing, the present value of the future payments necessary to service the debt will equal the tax reduction. However, the collectivity does not act as a unit, so a statement about aggregate equivalence is irrelevant to any effort to explain fiscal conduct. What matters for collective action is the direction of individual desires as these are mediated through political and fiscal institutions. For instance, people in higher age ranges will find debt to be less costly than taxation, increasingly so with age. Compare a tax of $1,000 now with a perpetual debt that entails payments of $100 when the appropriate discount rate is ten per cent. In terms of perpetuity, the debt and the tax are equivalent. For a younger person who might look forward to 50 taxpaying years, the present value of the debt is $991. For an older person who might only have ten years of tax-paying life expectancy left, the present value of the debt is but $614.

To be sure, it could be claimed that the older person has some bequest motivation towards heirs. If so, that older person would treat the debt obligation as continuing beyond his life. But not all older people have heirs. And of those that do, not all of them seem to have the types of bequest motives that generate Ricardian equivalence. This point gets to another significant feature of Buchanan's thought: his unwillingness to make statements based on aggregates without exploring the underlying structural patterns to which those aggregates pertain. After all, aggregates are not entities that act, and in Buchanan's approach collective action must be generated out of choices by discernible, acting individuals, as these choices are mediated through institutional frameworks for making collective choices.

The literature on public choice has, of course, exploded since 1967, with entrées to this literature provided by such compendia as Mueller (1997), Rowley and Schneider (2004), and Shughart and Razzolini (2001). A good deal of that literature has carried forward the effort of Buchanan and his Italian forebears to articulate the impact of political institutions on collective outcomes.

From Wicksell to Constitutional Political Economy

Where public choice examines the impact of political and fiscal institutions on collective outcomes, constitutional political economy examines the impact of constitutional rules on post-constitutional outcomes. The seminal statement of constitutional political economy is the *Calculus of Consent* (co-authored with Gordon Tullock), which the authors described as simply an elaboration with economic logic of the American constitutional framework of 1789. According to that framework, government is established by the consent of the governed, which provides unanimity as the conceptual starting point, just as it did for Wicksell (Wagner 1988, explores the relationship between Wicksell and the *Calculus of Consent*). While unanimity is the conceptual starting point, any effort actually to implement unanimity will confront free riders and strategic hold-outs. If everyone's consent is required to undertake collective action, some people will be tempted to withhold their consent, not because they object to the action but because they are acting strategically to shift the fiscal terms of the action in their favour. Such strategic efforts at securing distributional gain can sabotage projects that are genuinely beneficial to all. Consequently, people may

reasonably agree to be bound by something less than unanimous consent.

Buchanan and Tullock conceptualized a trade-off between decision costs and external costs, as these are viewed from the perspective of participants in collective choice. Decision costs are the costs people bear in trying to reach a collective decision. The greater the degree of consent required, the higher will be those costs due to such things as free riding and strategic bargaining. External costs are the costs that individuals bear when collective choices run contrary to their desires. These costs will fall with increases in the degree of consent required to take to collective action, and will vanish when unanimity is required. An optimal voting rule, formally speaking, will result when the sum of those costs is minimized. With this analytical construction, Buchanan and Tullock provided a rationalization for Knut Wicksell's support for some super-majority rule within a parliamentary assembly, as illustrated by references to three-quarters and four-fifths consent.

A voting rule is a simple scalar. Actual constitutional frameworks for collective choice contain a vector of characteristics, and to some extent those other characteristics can substitute for greater inclusivity in the degree of consent required. For instance, a representative assembly that is bicameral can achieve a greater degree of consensus with a less inclusive voting rule than would be possible within a unicameral assembly. Legislative action, moreover, can be filtered in various fashions through different parliamentary rules. There are many margins along which political and fiscal institutions can be modified, and with post-constitutional politics adapting to whatever constitutional framework is in place.

There are two levels of analysis in Buchanan's analytical schema: constitutional and post-constitutional. Post-constitutional politics, public choice, represents the working out of interactions among political participants within the context of some particular institutional arrangement. Constitutional politics concerns the selection among possible institutional arrangements. Buchanan's distinction between constitutional and post-constitutional politics calls forth the distinction between choosing the rules of a game and choosing strategies by which to play a game. For Buchanan, reform is a constitutional and not a post-constitutional matter.

Consider, for instance, his approach to progressive income taxation. Where the Anglo-Saxon sacrifice theorists sought to specify the degree of progressivity that some exogenous authority should impose on a society, Buchanan sought to probe the circumstances under which people might choose to employ progressivity in taxing themselves. In several places, he explores the conditions under which people might support progressive income taxation as a form of income insurance. Progressive taxation, as compared with proportional taxation, allows people to achieve some smoothing of consumption in the presence of fluctuating income. The purchase of insurance, after all, is a constitutional and not a post-constitutional activity: people purchase insurance before they have had accidents and not after. To the extent that such formulations have merit, what appears to be redistribution when seen from an *ex post* perspective might represent mutual gains from trade when viewed from an *ex ante*, constitutional perspective.

Alternatively, consider the treatment of broad-based taxation in Buchanan and Congleton (1998). Without a constitutional requirement of uniformity in taxation, post-constitutional politics will generate increasingly complex revenue systems as tax favours are granted or removed within the political marketplace. While the resulting narrowing of the tax base imposes excess burdens on market participants, it also warps processes of collective choice. For instance, those who are favoured by the resulting fiscal discrimination will support more collective activity than they would otherwise. With the continual churning of the tax code that results, however, most participants may end up worse off than they would have been under a simple system of tax uniformity.

Buchanan's Legacy

Until the late 1930s there was a flourishing Continental orientation towards public finance that

stood in contrast to the Anglo-Saxon orientation, and pretty much along the lines articulated by Buchanan in *Public Finance in Democratic Process* (this thesis is elaborated in Backhaus and Wagner 2005). Within this orientation, public finance was a multidisciplinary field of study, with a home in economics but with tentacles that reached out into such fields as politics, law, and public administration. Buchanan has carried forward the Continental approach to public finance, and has given it new life through his many creative works.

See Also

- ▶ Constitutions, Economic Approach to
- ▶ Sovereign Debt

Selected Works

1958. *Public principles of public debt.* Homewood: Richard D. Irwin.

1962. (With G. Tullock.) *The calculus of consent.* Ann Arbor: University of Michigan Press.

1967. *Public finance in democratic process.* Chapel Hill: University of North Carolina Press.

1977. (With R. Wagner.) *Democracy in deficit.* New York: Academic Press.

1992. *Better than plowing.* Chicago: University of Chicago Press.

1998. (With R. Congleton.) *Politics by principle, not interest.* Cambridge: Cambridge University Press.

1999–2002. *The collected works of James M. Buchanan.* 20 vols. Indianapolis: Liberty Fund. (This collection contains the preponderance of his academic work, both books and papers, through the late 1990s.).

Bibliography

Backhaus, J., and R. Wagner. 2005. From continental public finance to public choice: Mapping continuity. *History of Political Economy* 37(supplement): 314–332.

Mueller, D., eds. 1997. *Perspectives on public choice.* Cambridge: Cambridge University Press.

Musgrave, R., and A. Peacock, eds. 1958. *Classics in the theory of public finance.* London: Macmillan.

Puviani, A. 1960. *Die Illusionen in der öffentlichen Finanzwirtschaft.* Berlin: Dunker & Humblot . [A German translation of A. Puviani, *Teoria della illusione finanziaria,* 1903.].

Rowley, C., and F. Schneider. 2004. *The encyclopedia of public choice.* Dordrecht: Kluwer Academic Publishers.

Shughart, W. II, and L. Razzolini, eds. 2001. *The Elgar companion to public choice.* Cheltenham: Edward Elgar.

Wagner, R. 2003. Public choice and the diffusion of classic Italian public finance. *Il pensiero economico* 11: 271–282.

———. 1988. *The calculus of consent:* A Wicksellian retrospective. *Public Choice* 56: 153–166.

Bücher, Karl Wilhelm (1847–1930)

Bertram Schefold

Keywords

Bücher, K. W.; Division of labour; Exchange; German Historical School; Gifts; Industrial organization; Law of mass production; Menger, C.; Methodenstreit; Rhythm; Rodbertus, J. K.; Schmoller, G. von; Stages theory of economic development

JEL Classifications

B31

Karl Bücher was born in Kirberg (Germany) into a poor family. He studied history and classical philology in Bonn and Göttigen. Bücher first worked as a journalist for the liberal *Frankfurter Zeitung,* and from 1881 taught political economy in Dorpat, Basle, Karlsruhe and Leipzig, where he retired in 1917.

Bücher is counted among the outstanding economists of the German 'younger' historical school. He remained, however, independent in his economic thinking. He did not adhere to the inductive

method and in the Methodenstreit he sided with Menger against Schmoller. Although he advocated the adoption of social policy measures by the state, he confessed to being a liberal and did not follow the protectionist and state interventionist line of the 'Kathedersozialisten' (socialists of the chair). An important contribution to economics was Bücher's 'law of mass production', which described the relationship between production costs and output in industrial manufacturing. Moreover, Bücher carefully analysed the organization of the labour process and the division of labour (1893, pp. 261–334). His study on the importance of rhythm for the working process in pre-industrial societies is extremely interesting and may be regarded as his most original work (1896). He described how workers transformed monotonous physical labour through the adoption of rhythmic repetitions of their movements. By adjusting the work speed to this rhythm, the working process was both eased and intensified. Such a rhythm could be generated, for example, by singing. Bücher gave vivid examples of typical work songs and particularly described the role played by work songs in combining large masses of workers to carry out large-scale works. However, a precondition for all this was the worker controlling his individual work speed and dominating his working instruments. The fact that in modern industry this was no more the case led Bücher to interesting reflections on man and work in our industrial environment (1896, pp. 112–117).

Bücher's historical research focused on primitive people, antiquity and the Middle Ages. His analysis of primitive people (1893, pp. 1–82; 1918, pp. 1–26) was too generalized and did not grasp fully the extreme complexity of economic relations among these peoples. However, in his elaborate research on the distinction between exchange and gift he anticipated some of the problems which modern ethnology would later discuss. His studies on the economies of ancient Rome and Greece were important because they contributed to the refutation of authors who described these economies as simply capitalistic. Among his contributions on the Middle Ages were studies on the social situation of women and journeymen, and a demographic study on medieval Frankfurt, where Bücher applied statistical methods (1886; 1922).

Bücher developed a theory of stages of economic development (1893, esp. pp. 83–160), where he distinguished between the household economy (Hauswirtschaft) of classical antiquity (in accordance with J.K. Rodbertus' notion of the *oikos* economy), the town economy (Stadtwirtschaft) of the Middle Ages, and the national economy (Volkswirtschaft), that is, the extensive exchange economy of modern times. The role of exchange served as the central distinctive criterion: exchange was supposed to be virtually absent in the household economy, which is the reason why the characterization of antiquity (where trade had been more important than Bücher thought) as a household economy was inaccurate. Exchange was confined to locally produced commodities and local markets in the medieval town economy, and dominating every sphere of economic life in the 'national economy'.

Bücher may also be regarded as one of the founders of journalism as an academic discipline. He especially focused on the role of the press for public opinion and the problems raised by the capitalist and profit-oriented structure of the press.

Selected Works

1886. *Die Bevölkerung von Frankfurt am Main im XIV. und XV. Jahrhundert: Socialstatistische Studien.* Tübingen: Laupp.

1893. *Die Entstehung der Volkswirtschaft.* 1. Sammlung. 16th ed. Tübingen: Laupp, 1922. Trans. as *Industrial evolution*. Toronto: University of Toronto Press, 1901.

1896. *Arbeit und Rhythmus.* Abhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaft. Leipzig: Hirzel, 1897.

1918. *Die Entstehung der Volkswirtschaft.* 2. Sammlung. 8th ed. Tübingen: Laupp, 1925.

1922. *Beiträge zur Wirtschaftsgeschichte.* Tübingen: Laupp.

Buckle, Henry Thomas (1821–1862)

F. Y. Edgeworth

Buckle led a student's recluse life, devoted to the great historic work which he left unfinished on his death in his 41st year (1862). In the introduction to this work, the principle that human actions obey laws verifiable by statistics, was, as Mill says (*Logic*, bk. vi, chapter xi, § 1), 'most clearly and triumphantly brought out' by Buckle. Mill does not however agree in the opinion 'that the moral qualities of mankind are little capable of being improved,' and conduce little [to] the progress of society (ibid., § 2). Dr. Venn has protested more strongly against Buckle's fatalistic interpretation of statistics (*Logic of Chance*, 2nd edn, pp. 235–241). An erroneous impression of the futility of human effort is conveyed by such statements as 'suicide is merely a product of the general condition of society, and the individual felon only carries into effect what is a necessary consequence of preceding circumstances' (Venn, *Logic of Chance*, chapter xviii, § 14; Buckle, *History of Civilisation*, vol. i, chapter i). The same disposition to underrate the force of human will appears in Buckle's theories as to the influence of physical conditions on wages and population: 'There is a strong and constant tendency in hot countries for wages to be low, in cold countries for them to be high. The evil condition of Ireland was the natural result of cheap and abundant food' (*History of Civilisation*, chapter ii). He here maintains that 'potato philosophy of wages', which Walker stigmatized (*Political Economy*, bk. v, chapter iii). Buckle's economical reflections are indeed not always sound, but they bear the impress of originality, enhanced by copious learning and recondite references. His account of the discoveries made by political economists is masterly (chapter iv). The remarks on the leading economists, in particular Adam Smith and Hume, are instructive, even when disputable. The description of Adam Smith's method as *deductive*, is a half-truth characteristic of Buckle.

Selected Works

1857–61. *History of civilisation in England*, 2 vols. London.
1872. In *Miscellaneous and posthumous works*, ed. H. Taylor. London.

References

Huth, A.H. 1880. *The life and writings of Henry Thomas Buckle*. London.
Mill, J.S. 1843. *A system of logic*. London.
Stuart-Glennie, J.S. 1875. *Pilgrim memories, or travel and discussion in the birth-countries of Christianity with the late Henry Thomas Buckle*. London.
Venn, J. 1866. *The logic of chance*. London.
Venn, J. 1881. *Symbolic logic*. London.
Venn, J. 1889. *The principles of empirical or inductive logic*. London.
Walker, F.A. 1883. *Political economy*. New York/London.

Budget Deficits

William G. Gale

Abstract

This article describes alternative measures of the federal budget deficit, discusses traditional and non-traditional channels through which deficits can affect the economy, and summarizes research on the effects of deficits on national saving and interest rates.

Keywords

Aggregate consumption; Budget deficits; Consumption function; Crowding out; Euler equations; Financial market expectations; Fiscal policy; Generational accounting; Interest rates; Labour supply; National income;

National saving; Public debt; Ricardian equivalence theorem; Ricardo, D; Small open economy view

JEL Classifications
H4

Federal budget deficits reflect the extent to which current federal spending policies are not being financed with current federal tax policies, and can have significant effects on national saving and interest rates.

Economists have explored the effects of budget deficits extensively, and analysis of the aggregate effects of fiscal policy dates back at least to the work of David Ricardo. Modern academic interest was reinvigorated by the work of Barro (1974) and others, and by the large US federal budget deficits in the 1980s and early 1990s. These factors led to a substantial amount of research that is summarized in several excellent surveys (Barro 1989; Barth et al. 1991; Bernheim 1987, 1989; Elmendorf and Mankiw 1999; Seater 1993). The rapid but short-lived transition to unified budget surpluses in the late 1990s, followed by the sharp reversal in budget outcomes since 2000, has raised interest in this question again.

The budget deficit can be defined in many different ways, and the most appropriate measure is likely to depend on the particular model or application of interest. For any measure of the deficit, which is a flow during a given time period, there is an analogous measure of the public debt, which is a stock at a given point in time and which represents the net accumulation of the associated deficits over all previous time periods.

The most widely used measure of the US federal deficit – the unified budget balance – is fundamentally (but not exactly) a cash-flow metric that includes both the Social Security and non-Social Security components of the federal budget. In a first approximation, the unified deficit shows the extent to which the government borrows or lends in credit markets. For some purposes, it is more informative to examine the primary budget, which excludes interest payments on the public debt (that is, it is equal to the unified budget balance minus net interest payments). The standardized budget balance adjusts the unified budget for the business cycle and special items. All these measures share a basic focus on cash flow.

Broader measures of the budget deficit look beyond cash flow and take into account the implicit or explicit promises embedded in current government policies, even if such promises do not result in current-period cash flow. Generational accounting, for example, aims to tally the net debt that each generation or birth cohort faces (see Auerbach et al. 1991 for discussion of generational accounting and Auerbach et al. 2003 for discussion of alternative measures of the deficit). However, it is unclear how the market and households value implicit debts relative to the government's explicit debt. Thus, while the importance of the broader measures is clear conceptually, this article focuses mostly on the cash-flow related measures of the deficit.

In the fiscal year 2005, the unified US federal deficit was about 2.6% of the GDP, and the standardized deficit was about 1.8% (Congressional Budget Office 2006). The current budget situation would largely not be a concern if future fiscal prospects were auspicious. Unfortunately, the longer-term budget outlook is dismal, primarily because of projected rising expenditures on health care and programmes for the elderly (Congressional Budget Office 2005).

Economic Effects of Budget Deficits: Traditional Channels

Economists tend to view the aggregate effects of tax cuts from one of three perspectives. To sharpen the distinctions, consider deficits induced by changes in the timing of lump-sum taxes, with the path of government purchases and marginal tax rates held constant. Under the Ricardian equivalence hypothesis, such deficits are fully offset by increases in private saving and have no effect on national saving, interest rates, exchange rates, future domestic production, or future national income. A second model, the small open economy view, suggests that budget deficits

reduce national saving, but induce increased international capital inflows that finance the entire reduction in national saving. As a result, domestic production does not decline and interest rates do not rise, but future national income falls because of the burden of repaying the increased borrowing from abroad. A third model, which we call the conventional view, suggests that deficits reduce national saving and that the reduction in national saving is at least partly reflected in lower domestic investment. In this model, budget deficits partly crowd out private investment and partly increase borrowing from abroad; the combined effect reduces future national income and future domestic production. The reduction in domestic investment in this model is facilitated by an increase in interest rates, establishing a connection between deficits and interest rates.

It is worth emphasizing that the relationship between deficits and national saving is central to analysis of the economic effects of fiscal policy. National saving, which is the sum of private and government saving, finances national investment, which is the sum of domestic investment and net foreign investment. Higher national saving raises the capital stock owned by the nation's citizens and thus raises future national income.

An increase in the budget deficit reduces national saving unless it is fully offset by an increase in private saving. If national saving falls, national investment and future national income must fall as well, if all else remains equal. Therefore, to the extent that budget deficits reduce national saving, they reduce future national income. This reduction in future national income occurs even if there is no increase in domestic interest rates. In the case where there is no rise in domestic interest rates, the reduction in national saving associated with budget deficits would manifest itself solely in increased borrowing from abroad (as under the small open economy view). This is the sense in which the effect of deficits on interest rates and exchange rates (the distinction between the small open economy view and the conventional one) is subsidiary to the question of the effects on national saving (the Ricardian view versus the other two).

A key consideration is that the results above consider only the effects of increased budget deficits or debt per se. A full analysis of the effects of public policies on economic growth should take into account not only the effects of increased deficits and debt but also the direct effects of the spending programmes or tax reductions that cause them. The effects of fiscal policies on both economic performance and interest rates depend not only on the deficit but also on the specific elements of the policies generating that deficit. For example, spending one dollar on public investment projects would increase the unified budget deficit by one dollar, but the net effect on future income would depend on whether the return on the public investment project exceeded the return on the private capital that would have instead been financed by the national saving crowded out by the deficit. Similarly, a deficit of one per cent of GDP caused by reducing marginal tax rates will generally have different implications for both national income and interest rates from a deficit of one per cent of GDP caused by increasing government purchases of goods and services.

Economic Effects of Budget Deficits: Non-traditional Channels

Beyond their direct effect on national saving, future national income and interest rates, deficits can affect the economy in other ways. For example, increased deficits may cause investors gradually to lose confidence in national economic stability and leadership. As Truman (2001) emphasizes, a substantial fiscal deterioration over the longer term may cause 'a loss of confidence in the orientation of US economic policies'. Such a loss in confidence could then put upward pressure on domestic interest rates, as investors demand a higher risk premium on dollar-denominated assets. The costs of current account deficits – which are in part induced by large budget deficits – may even extend beyond narrow economic ones. More broadly, Friedman (1988, p, 76) notes that 'World power and influence have historically accrued to creditor countries. It is not coincidental that America emerged as a world

power simultaneously with our transition from a debtor nation . . . to a creditor supplying investment capital to the rest of the world.'

Both the traditional models and the non-traditional effects noted above focus on gradual negative effects from reduced national saving. This focus may be too limited, however, in that it ignores the possibility of much more sudden and severe adverse consequences. In particular, the traditional analysis of budget deficits in advanced economies does not seriously entertain the possibility of explicit default or implicit default through high inflation. If market expectations regarding the probability of default were to change and investors had difficulty seeing how the policy process could avoid extreme steps, the consequences could be much more sudden and severe than traditional estimates suggest. The role of financial market expectations in this type of scenario is central. One of the key triggers would occur if investors begin to doubt whether the strong historical commitment to avoiding substantial inflation would be weakened in order to reduce the real value of the public debt (Ball and Mankiw 1995; Rubin et al. 2004).

Although this article does not explicitly incorporate non-traditional effects into the discussion below, such effects serve as an important reminder of why budget deficits, especially chronic deficits, could exert large adverse effects on US economic performance. The focus on traditional effects is certainly justifiable in the context of historical analysis of post-war data from the United States. That does not imply, however, that to ignore such issues is appropriate when examining the likely impacts of future deficits. The nation has never before faced substantial deficits that are projected to be sustained and indeed to grow over many decades.

Deficits and Consumption

Testing the effect of deficits on aggregate consumption, with government spending held constant, is an important focus of analysis for several reasons. First, these analyses provide a direct test of whether the timing of tax collections affects the economy, with other factors controlled for. Second, the aggregate time series tests measure the *magnitude* of the effects in question. This is particularly important because virtually no one claims that Ricardian equivalence is literally true. Rather, the controversy is over the extent to which Ricardian equivalence is a good approximation of the aggregate impact of fiscal policies.

There is a wide variety of research findings from studies of aggregate consumption and fiscal policy, in part because of a variety of difficult econometric issues. Barro (1989) and Elmendorf and Mankiw (1999) conclude that the literature is inconclusive. Seater (1993) concludes that, once the studies are corrected for econometric problems, Ricardian equivalence is corroborated – or at least that it is not possible to reject Ricardian equivalence. Bernheim (1989) concludes that, once the studies are normalized appropriately, Ricardian equivalence should be rejected.

One strand of the literature specifies consumption functions and then tests for the effects of fiscal policy. Perhaps the best-known study in this area is Kormendi (1983), who finds no evidence of non-Ricardian effects. This work has spawned significant research, including three sets of exchanges in the *American Economic Review*. Recent research, however, has extended the Kormendi results in three ways: using more recent data, which captures significant variation in budget outcomes; controlling for measures of marginal tax rates; and (in the United States) allowing federal and state fiscal variables to have different effects on consumption. The last issue is particularly relevant because the states collect a significant share of their revenue through consumption taxes, which would be expected to vary *positively* with consumption, whereas other taxes would be expected, at least in non-Ricardian theory, to vary negatively. With these extensions, the results suggest that about 30–46 cents of every dollar in federal tax cuts is spent in the same year (Gale and Orszag 2004). This is a rejection of the Ricardian view.

Another strand of the literature focuses on Euler equation tests (relating to the growth rate of consumption, as opposed to the tests above,

which examine consumption levels), with mixed results. As Bernheim (1987) points out, Ricardian equivalence can fail even if the Euler equation does not, and vice versa. Nevertheless, some studies have found substantial effects of fiscal policy on consumption using the Euler framework, most recently Gale and Orszag (2004), who find that about 50–85 cents of every dollar in tax cuts is spent in the first year, with most of the effects measured precisely. This range is consistent with some previous assessments, but it is inconsistent with the Ricardian prediction of a full offset from private saving.

Deficits and Interest Rates

The effects of fiscal policy on interest rates have also proven difficult to pin down statistically. The issues include the appropriate definition of deficits and debt, whether deficits or debt should be the variable of interest, the difficulty of distinguishing expected and unexpected changes, and the potential endogeneity of many of the key explanatory variables (see Bernheim 1987; Elmendorf and Mankiw 1999; Seater 1993).

In part because of these statistical issues, the evidence from the empirical literature as a whole is mixed. However, the key role of *expected* deficits rather than current deficits is sometimes overlooked. As Feldstein (1986, p. 14) has written, 'it is wrong to relate the rate of interest to the concurrent budget deficit without taking into account the anticipated future deficits. It is significant that almost none of the past empirical analyses of the effect of deficits on interest rates makes any attempt to include a measure of expected future deficits.' Since financial markets are forward-looking, to exclude expectations could bias the analysis towards finding no relationship between interest rates and deficits. In fact, studies that incorporate more accurate information on expectations of *future* sustained deficits tend to find economically and statistically significant connections between anticipated deficits and current interest rates. Gale and Orszag (2004) show that, of the 19 papers incorporating timely information on projected deficits, 13 find predominantly positive, significant effects between anticipated deficits and current interest rates, five find mixed effects, and only one finds no effects. The other studies in the literature that find no significant effect are disproportionately those that do not take expectations into account at all or do so only indirectly through a vector autoregression. Thus, while the literature as a whole, taken at face value, generates mixed results, analyses that focus on the effects of anticipated deficits tend to find a positive and significant impact on interest rates.

The challenge in incorporating market expectations about future deficits is that such expectations are not directly observable. An important caveat to the whole literature, then, is that, to the extent that proxies for expected deficits are imperfect reflections of current expectations, the coefficient on the projected deficit will tend to be biased towards zero because of classical measurement error, and the studies would tend to *underestimate* the effects of deficits on interest rates.

Even among studies that use expected deficits, one potential concern is that the business cycle could be affecting current yields. Laubach (2003) suggests a novel way to resolve this issue: he examines the relationship between projected deficits (or debt) and the level of real *forward* (five-year ahead) long-term interest rates. The underlying notion is that current business cycle conditions should not influence the long-term rates expected to prevail beginning 5 years ahead. Laubach uses projections of the US Congressional Budget Office and Office of Management and Budget, and finds that a one percentage point increase in the five-year-ahead projected deficit-to-GDP ratio raises the five-year-ahead ten-year interest rate by between 24 and 40 basis points, and that a one percentage point in the projected debt-to-GDP ratio raises the long-term forward rate by between 3.5 and 5.5 basis points. The deficit-based results are not dissimilar from the debt-based results. Consider, for example, an increase in the budget deficit equal to one per cent of GDP in each year over the next 10 years. After 10 years, that would raise government debt by roughly ten per cent of GDP. The deficit-based results in Laubach would suggest about a 30 basis point increase in interest

rates, whereas the debt-based results would suggest about a 45 basis point increase.

Using a similar framework, Engen and Hubbard (2004) obtain somewhat smaller effects while Gale and Orszag (2004) obtain somewhat larger effects. Indeed, despite a rancorous public debate, there appears to be a surprising degree of convergence in recent estimates of the effects of fiscal policy on interest rates, with a variety of econometric studies implying that a sustained one per cent of GDP increase in unified deficits over 10 years would raise interest rates by 30–60 basis points. The relationship between deficits and interest rates not only provides further evidence against the Ricardian view, but also implies that the conventional view is a better description of reality for the United States than the small open economy view. Ardagna et al. (2004) find even stronger results in a panel of 16 Organisation for Economic Co-operation and Development (OECD) countries over several decades.

Conclusion

Sustained federal budget deficits have two sets of effects. The direct effect of the increase in government borrowing is to reduce national saving and raise long-term interest rates, often by empirically sizable amounts. The other set of effects depends on the specific tax or spending policies that were chosen to create the deficits. These findings have significant implications. First, both the consumption and the interest rate results reject the Ricardian view of the world. Second, the interest rate results reject the small open economy view, at least as it applies to the US economy. Third, the results suggest that the sustained deficits facing the nation will impose significant economic costs. Fourth, some tax-cut policies that have traditionally been considered growth-enhancing may actually backfire, because the generally positive effect of the tax rate cut on labour supply and investment, if interest rates are held constant, can be offset by the impact of the deficit on interest rates and on national saving. While it would be wrong to conclude that all these issues are decisively resolved in the economics literature, there is more than strong enough evidence to raise concerns about sustained projected future deficits.

See Also

- ▶ Crowding Out
- ▶ New Open Economy Macroeconomics
- ▶ Ricardian Equivalence Theorem

Bibliography

Ardagna, S., F. Caselli, and T. Lane. 2004. *Fiscal discipline and the cost of public debt service: Some estimates for OECD countries*, Working paper No. 10788. Cambridge, MA: NBER.

Auerbach, A.J., J. Gokhale, and L.J. Kotlikoff. 1991. Generational accounts: A meaningful alternative to deficit accounting. In *Tax policy and the economy*, ed. D. Bradford. Cambridge, MA: NBER.

Auerbach, A.J., W.G. Gale, P.R. Orszag, and S.R. Potter. 2003. Budget blues: The fiscal outlook and options for reform. In *Agenda for the nation*, ed. H.J. Aaron, J.M. Lindsey, and P.S. Nivola. Washington, DC: Brookings Institution.

Ball, L., and N.G. Mankiw. 1995. What do budget deficits do? In *Budget deficits and debt: Issues and options*. Kansas City: Federal Reserve Bank of Kansas City.

Barro, R.J. 1974. Are government bonds net worth? *Journal of Political Economy* 82: 1095–1117.

Barro, R.J. 1989. The Ricardian approach to budget deficits. *Journal of Economic Perspectives* 3(2): 37–54.

Barth, J.R., G. Iden, F.S. Russek, and M. Wohar. 1991. The effects of federal budget deficits on interest rates and the composition of domestic output. In *The great fiscal experiment*, ed. R.G. Penner. Washington, DC: Urban Institute Press.

Bernheim, B.D. 1987. Ricardian equivalence: An evaluation of theory and evidence. *NBER Macroeconomics Annual* 2: 263–304.

Bernheim, B.D. 1989. A neoclassical perspective on budget deficits. *Journal of Economic Perspectives* 3(2): 55–72.

Congressional Budget Office. 2005. *The long-term budget outlook*. Washington, DC: Congressional Budget Office.

Congressional Budget Office. 2006. *The budget and economic outlook: Fiscal years 2007 to 2016*. Washington, DC: Congressional Budget Office.

Elmendorf, D.W., and N.G. Mankiw. 1999. Government debt. In *Handbook of macroeconomics*, vol. 1C, ed. J.-B. Taylor and M. Woodford. Amsterdam: North-Holland.

Engen, E.M., and R.G. Hubbard. 2004. Federal government debt and interest rates. In *NBER macroeconomic annual 2004*, ed. G. Mark and K.S. Rogoff. Cambridge, MA: MIT Press.

Feldstein, M.S. 1986. *Budget deficits, tax rules, and real interest rates*, Working paper No. 1970. Cambridge, MA: NBER.
Friedman, B. 1988. *Day of reckoning: The consequences of American economic policy under Reagan and after*. New York: Random House.
Gale, W.G., and P.R. Orszag. 2004. Budget deficits, national saving, and interest rates. *Brookings Papers on Economic Activity* 2004(2): 101–210.
Kormendi, R. 1983. Government debt, government spending, and private sector Behavior. *American Economic Review* 73: 994–1010.
Laubach, T. 2003. *New evidence on the interest rate effects of budget deficits and debt*, Finance and economics discussion series No. 2003–12. Washington, D.C: Board of Governors of the Federal Reserve System.
Rubin, R.E., P.R. Orszag, and A. Sinai. 2004. *Sustained budget deficits: Longer-run U.S. economic performance and the risk of financial and fiscal disarray.* Paper presented to the AEA-NAEFA Joint Session, Allied Social Science Associations Annual Meetings, 4 January, San Diego.
Seater, J.J. 1993. Ricardian equivalence. *Journal of Economic Literature* 31: 142–190.
Truman, E.M. 2001. *The international implications of paying down the debt*, Policy brief 01–7. Washington, DC: Institute for International Economics.

Budget Projections

Douglas Holtz-Eakin

Abstract

This article surveys different approaches to the construction of government budget projections, illustrated with procedures from the United States Office of Management and Budget (OMB) and Congressional Budget Office (CBO). It sets out the several distinct steps that are required in budget projections, from macroeconomic forecasting to comparing projections with outcomes and analysing the sources of deviations.

Keywords

Aggregate demand; Budget projections; Budget window; Business cycles; Forecasting; Intertemporal incentives; Scoring; Uncertainty

JEL Classifications

H4

Budget projections are central to governmental policymaking. In general, budgeting is the practice of devoting economic resources to policy objectives and providing specific means for raising these resources. A typical budget process includes budget proposals, review, adoption, and execution. Budget projections inform the process by providing estimated values for government revenues, government spending, and other budgetary concepts over a specific planning horizon (often referred to as the 'budget window'). Budgetary projections are made under specific assumptions, for differing government programmes, using alternative approaches as part of the budgetary process. We discuss each in turn, with examples drawn from the United States federal government.

Threshold assumptions for budget projections fall along two dimensions: economic and policy.

Economic Assumptions

One approach to developing a budget projection is based on a comprehensive economic forecast, inclusive of any possible future business cycle fluctuations. In this instance, the result is a projection of the potential future outlays, receipts, and budget deficit or surplus. In the United States, both the White House Office of Management and Budget (OMB) and the Congressional Budget Office (CBO) adopt a variant of this approach in which the near-term forecast incorporates the state of the business cycle, while projections beyond the first two years assume an average of full employment.

Alternatively, it is sometimes assumed that the economy operates continuously at full resource utilization with no cyclical fluctuations. In this instance, the budget projections are often referred to as 'cyclically adjusted' or 'full-employment' projections of the budget and its balance.

Each approach serves distinct purposes. Budget projections are necessary, for example, to

anticipate the cash-flow borrowing needs of the government on a year- by-year basis. In contrast, cyclically adjusted budget projections are useful for judging whether current deficits or surpluses are reflective of the state of the economy, and thus the degree to which fiscal policies are sustainable over the longer term.

Policy Assumptions

The future path of the budget also depends on the evolution of tax and spending policies. In constructing the budget projection, one possible assumption is that current policies (or current laws) remain unchanged. Such a projection – known alternatively as a budget baseline projection or current services projection – provides a means by which to judge the future implications of current policies and a benchmark (or baseline) against which to measure the impact of policy changes.

Two issues arise in constructing and interpreting baseline budget projections. The first is the rules for anticipating any necessary future policy actions. For example, in the US federal budget a large fraction (roughly two-thirds in 2007) of spending results from 'mandatory' (or 'direct') spending programmes in which laws authorize automatic expenditures to eligible parties. Common examples are Social Security, Medicare, and farm support programmes. In these instances, projections of spending rely on combining rules of the programmes with projections of eligible populations and their relevant characteristics. An issue arises when the legal authorization for a programme expires during the projection period, requiring an assumption regarding whether spending will stop entirely or continue as if the current programme remains in place. (In the United States, 'large' programmes – spending in excess of $50 million – are assumed to continue.)

The remainder of spending (over one-third in 2007) is 'discretionary' and determined by the annual decisions of Congress. Consistent with the spirit of projecting current policy, baseline projections typically assume that this type of spending continues (in real, inflation-adjusted terms) exactly as in the most recently completed budget. An implication of this procedure is that baseline projections of discretionary spending may be heavily influenced by transitory policy events such as emergency spending.

These types of swings in projected spending are illustrative of the second key feature of baseline or current services projections. These projections are not forecasts of actual budget outcomes, but rather tools to inform the budgetary process.

A second approach is to embed in the budget projections a specific path for future policies, that is, to construct a policy-based budget projection. For example, the annual Presidential budget submitted to the US Congress is constructed under the assumption that all the proposed policies are adopted as requested. As with baseline budget projections, policy projections are not forecasts of actual budgetary outcomes.

Scoring

A topic closely related to budget projections is 'scoring' – the evaluation of the budgetary implications of policy proposals. Mechanically, scoring represents the difference between a policy-based projection and a baseline projection, thereby revealing the budgetary difference as a result of the specific policies.

Scoring budgetary proposals permits comparisons of alternative proposals on a consistent basis. Traditionally, scores have been constructed under the assumption that overall macroeconomic performance is unchanged by the policy proposal ('static scoring'). There are some proposals, however, of sufficient magnitude and impact on incentives (for example, tax reform) that it would be desirable to incorporate not only the direct budgetary impacts but also the budgetary feedbacks from changes in the overall levels of economic output and incomes ('dynamic scoring'). Incorporating economic impacts, however, raises issues in maintaining consistency in scoring across

proposals and details of executing the analysis (see Congressional Budget Office 2002; Joint Committee on Taxation 2006).

Steps for Budgetary Projections

Official governmental budget projections from, for example, the OMB and the CBO, are sophisticated, detailed exercises that require several distinct steps.

1. *Project macroeconomic performance.* The budget projection is built upon a macroeconomic forecast, including the path for real and nominal gross domestic product (GDP), the future rates of unemployment, the path for prices and inflation, and the path of future interest rates and exchange rates. As part of anticipating the near-term position in the business cycle, it is necessary to forecast the components of aggregate demand – consumption, residential and business investment, government spending, and net exports – as well as the determinants of the potential for overall output, such as capital stocks, labour force, and technological progress. Because of the importance of tax revenues to the budgetary projections, the projection of national income is more important than in other settings, imposing the requirement for projecting labour compensation, taxable versus non-taxable compensation, corporate profits, dividends, interest payments, and non-corporate business income.
2. *Impute a distribution to macroeconomic aggregates.* In the United States, personal income tax is progressive and heavily skewed towards the upper part of the income distribution (with the top one-half of households paying nearly all the income tax). Accordingly, the distribution of wage and salary earnings (as well as other components of household income) among households has a large impact on the overall level of tax receipts. In these circumstances, the macroeconomic forecast must be combined with microeconomic data drawn from tax returns and population surveys to provide accurate projections.
3. *Impose programme rules on the macroeconomic and microeconomic data to project spending and revenues*. For example, the projections for population, labour force, and the unemployment rate yield forecasts of the number of unemployed individuals. When combined with unemployment insurance programme rules, the unemployment forecast yields a projection of outlays for the unemployment insurance programme. Similarly, the projection of wage income, dividend payments, interest payments, and capital gains, along with distributional information on each, may be combined with parameters of the tax code to produce projections of individual income tax receipts.

An important aspect of this step is the sophistication of incorporating responses to incentives in the projections. For example, if current law indicates that tax rates will rise in the next several years, it is likely that intertemporal incentives may shift forward some economic activity (for example, labour supply) and some tax-based planning behaviours (for example, realization of capital gains to obtain lower tax rates). It is desirable to incorporate these responses in the projection.

4. *Check for internal consistency.* In some circumstances, budget projections involve an element of simultaneity. For example, fiscal projections (spending and taxes) are necessary to forecast near-term aggregate demand, while actual outlays and tax receipts depend upon the employment and incomes generated by economic activity. Accordingly, it is desirable to check whether the budget totals are consistent with the economic projection.
5. *Compare projections with actual outcomes to improve projections*. The accuracy of budget projections is an obvious concern. Hence it is desirable to do a comparison of actual outcomes with past projections to identify systematic sources of error and opportunities for improvement. In addition, a second desirable

attribute of projections is their credibility, which is aided by a transparent process for revealing differences between actual and projected outcomes, and a systematic analysis of the sources of deviation.

Uncertainty and Valuation in Budget Projections

Uncertainty

Budgetary projections are fraught with uncertainty. At the most basic level, the future is literally unknowable, and budgetary projections will be affected by the future course of macroeconomic fluctuations, variations in inflation, the path of interest rates, and so forth. The degree to which projections are uncertain is important information to policymakers. One approach to revealing the scale of uncertainty is to undertake the budget projections in a series of scenarios (for example, 'base case', 'faster growth and higher inflation', and 'slower growth and lower inflation'). The difficulty then becomes choosing scenarios that are representative of the likely fluctuations to be experienced.

A more complete and formal approach is to conduct the entire projection in the context of a stochastic simulation methodology. In this approach, historical joint distributions are constructed for the key inputs to the projection (GDP growth, inflation, interest rates, wages, and so forth). Undertaking a large number of projections, each based on a 'draw' from the joint distribution, permits policymakers to be presented with the full distribution of potential outcomes over the budget horizon.

A second type of uncertainty is important for individual programmes. In some cases, government budget flows are contingent upon uncertain outcomes. A prominent example is agriculture programmes that provide funds only in the event of poor harvests due to drought or other adverse events. How should budget projections be constructed for such programmes? Choosing a single scenario will probably yield a projection in which the programmes either have a budget impact *every* year or in no year – neither of which is a sensible projection. A simple solution is to use the average (perhaps over a historical period) as the projected value of the budget impact of the programme, with the logic being that the projection is never precisely correct, but on average informative. As above, however, an alternative is to undertake formal stochastic simulations of the programme in question and use the expected value of the programme as the budget projection.

Valuation

The practice of budgetary projections (and scoring) raises issues in the correct valuation of budgetary transactions. In the main, the goal is to value government purchases using market prices (and thereby adhering as closely as possible to private-sector measure of marginal cost and marginal benefit). Similarly, tax collections and transfers to individuals and governments are measured in dollar values. However, difficulties can arise in the consistent application of these principles.

A notable example is the provision of insurance and insurance-like programmes by the government. Adhering to the principles of taxes and transfers, the projections of these programmes consist of the future tax receipts by the government and payments to individuals. Put differently, the budget projection consists of the future cash flows, perhaps summarized in an expected value form. Note, however, that this budgetary treatment may complicate comparisons with an equivalent programme – the direct purchase of an equivalent private-sector insurance product, where the private-sector entity will charge a risk premium.

Bibliography

CBO (Congressional Budget Office). 2002. CBO testimony: Federal budget estimating. Statement of Dan L. Crippen, Director, before the Committee on the Budget, U.S. House of Representatives, 2 May. Online. Available at https://www.cbo.gov/sites/default/files/107th-congress-2001-2002/reports/05-02-testimony.pdf. Accessed 14 Feb 2007.

Joint Committee on Taxation. 2006. *Exploring issues in the development of macroeconomic models for use in tax policy analysis* (JCX-19–06), 16 June. Online. Available at http://www.jct.gov/x-19-06.pdf. Accessed 14 Feb 2007.

Budgetary Policy

M. H. Peston

The subject of budgetary policy in the period following the *General Theory* concerns the impact of public expenditure and taxation on aggregate demand. More recently some attention has also been paid to the relationship between the budget and aggregate supply. In both cases emphasis must be placed on the word 'aggregate'. The structure of government expenditure and taxation will also have an impact which can be studied at the microeconomic level in terms of effects on individual firms, households and markets, but that is not what is normally covered by the present heading.

Public expenditure may be classified into two main components; expenditure on goods and services, and transfer payments. The former may be divided into capital and current expenditure, and the latter into capital and current transfers. Theoretically, this division may correspond to the economic distinction between 'using up' and 'adding to stock'. In practice the division is more broad brush, and a great deal of what would be recognized as an addition to the public sector's stock of capital is treated as current expenditure.

What is regarded in practice as a transfer payment is also somewhat blurred, and sometimes depends on a distinction between someone being employed by the public sector as opposed to being supported by it. Within the transfer heading most systems of national accounts differentiate the costs of servicing the national debt from the remainder.

The issue of debt itself, both long term and short term, is connected with the financing of a budgetary deficit. Some systems of national accounts treat the purchase of private sector financial assets as public expenditure (and their sale by the government as negative public expenditure). Others more properly regard such activities as akin to the issue and redemption of debt, and, thus, falling within the orbit of financial policy.

Turning to the taxation side, the most straightforward classification is into direct and indirect taxes. The former, comprising income and capital taxes, may be divided into taxes levied on firms and taxes levied on households. Property taxes, i.e. the rates, will also be included in this category. Indirect taxes include sales taxes, purchase taxes and VAT. Even though these are sometimes nominally levied on firms, there is a tendency in macroeconomics to assume that their incidence is such that they are actually levied on households. A similar point applies to corporation taxes which may be passed forward and should be treated as indirect taxes on households.

Given all that as background, the central macroeconomic propositions are as follows. An increase in government spending raises aggregate demand. The extent to which it does so depends on the form of the spending as mentioned above, and on the value of the multiplier. A decrease in taxation also varies aggregate demand, and again the scale of the effect depends on the tax in question and the multiplier. In this case, however, the process depends on the tax cuts raising disposable income and private spending.

To the extent that taxes are a function of income and expenditure, they will influence the size of the multiplier. The larger the marginal tax rates, the more any income or expenditure will leak into the government's coffers, and the lower the value of the multiplier.

Transfer payments, although they are classified as government expenditure, work via their effect on disposable income and private expenditure. Some transfer payments may be endogenous, e.g. unemployment and other social security payments vary inversely with income. (In the longer run the *rate* at which these and similar payments are paid is likely to vary directly with income.)

Because aggregate demand is an increasing function of government expenditure and a decreasing function of taxation, these instruments in all their complexity may be used to manipulate it. If aggregate demand is forecast to be too low compared with aggregate supply, and is not expected to adjust automatically and quickly, a combination of public expenditure increases and tax cuts may be used to improve the position.

A reversal of the instruments will deal with the case in which aggregate demand is excessive.

The government's budgetary position may be defined as the difference between tax revenues and expenditures. Because the various forms of expenditure and taxation have different effects on aggregate demand, the results of budgetary policy cannot be inferred simply from an examination of the budgetary position. It is necessary to look in detail at the budget to ascertain the net impact of fiscal action (*see* Balanced Budget Multiplier).

The totals of government expenditure and of taxation depend partly on the levels of national income and expenditure, i.e. they are endogenous. The multiplier is lower the more powerful these endogenous forces are. This means that exogenous shocks have smaller effects on the level of national income and employment. High marginal tax rates (and transfer payments) automatically cause expenditure to fall less as income falls. They are, therefore, called *automatic* or *built in stabilizers*. This automatic stability is not, of course, an unalloyed benefit. The economy may be stabilized well away from full employment, and endogenous leakages may make it very hard for increases in government or private expenditure to cause the economy to move in an appropriate direction.

The endogeneity of some government revenue and expenditure also complicates the interpretation of the budgetary position. Starting (say) from an initial position of budget balance, a surplus may result from an increase in private spending or a reduction in government spending. The former will cause national income and tax payments to rise. The latter will cause national income to fall. (It will also cause tax payments to fall but by less than the fall in government spending). It follows that the emergence of a surplus does not mean that policy intervention has been actively contractionary. Indeed, if national income has risen as a result of greater monetary ease, policy will actually have been expansionary.

It may also be inferred that the budget surplus or deficit must be examined in relation to the level of national income. As a first approximation, the change in this surplus relative to income, and suitably weighted by the different aggregate demand effects of its expenditure and tax components, will indicate the extent to which fiscal policy has become more or less expansionary.

Another way of approaching this sort of issue is to ask what the budgetary position would be at a constant state of national income. The typical normalization is that corresponding to full employment. As has been noted, the budget may move into deficit if exogenous forces cause national income to fall below full employment. Tax revenues will be less and transfer payments more. Increases in rates of tax and cuts in rates of transfer payment may remove the deficit, but only by lowering aggregate demand further still. It may then be relevant to note that the exogenous restoration of full employment would also restore budget balance or even give rise to an excess of revenue over expenditure. It is suggested, therefore, that as well as the actual budgetary position, it is useful to calculate the full employment surplus (or deficit). This could indicate more accurately whether the net effect of policy is expansionary or not.

The question next arises of the relationship between fiscal policy and other forms of macroeconomic intervention, notably monetary policy and incomes policy. On the former, if the budgetary position is not one of balance, there will be consequences connected with financing of the deficit or surplus. The government must borrow to finance its deficit, for example. It may do this in a way which increases the money supply. Alternatively, it may borrow long term from the non-bank private sector. It follows that there may be both flow wealth and liquidity effects causing private expenditure to change. To the extent that a deficit is financed in ways that make the private sector feel wealthier and more liquid, the direct expansionary fiscal effect will be accentuated. National income will rise further, and with tax revenue endogenous, eventually the deficit will disappear. It is also logically possible that the deficit is financed in ways which lower the private sector propensity to spend. Indeed, this may be strong enough to offset the original fiscal expansion. The deficit would then increase, and if this absolutely larger deficit continued to be financed in the same way, national income would go on contracting.

The connection between budgetary and financial policy is strengthened if the interest rate effects of the former are also taken into account. The interest on this year's national debt adds to required expenditure next year. An excessive deficit leading to a rise in the ratio of national debt to national income can in theory raise the ratio of interest payments to income, which *ceteris paribus* will increase indefinitely.

Turning to incomes policy, a larger fraction (usually more than half) of public expenditure consists of wages and salaries paid to public sector employees. Given the number of these workers, an increase in their pay adds to government expenditure and to aggregate demand.

Whether control of the public sector pay bill is regarded as budgetary policy or incomes policy may seem more a matter of terminology than of fundamental economics. But it must also be borne in mind that government demand for labour is important in many sections of the labour market, and what the government is willing to pay may also be used to forecast its policy intentions especially in regard to inflation.

For some purposes of analysis it is important to place budgetary policy in an international context. The expansionary effect of an increase in public expenditure (especially if it is tax financed) will depend on the government's propensity to import compared with the private sector's, as will the change in the balance of payments or the exchange rate.

Possibly more important than that, the effectiveness of an increase in government expenditure depends on whether the exchange rate is fixed or variable, and the degree to which capital is mobile internationally.

On the former, starting from less than full employment, an increase in government expenditure will worsen the trade balance. If the nominal exchange rate is allowed to fall, and domestic prices do not rise, so that there is also a real devaluation of the currency, exports will rise relative to imports. In other words, in standard Keynesian terms, fiscal policy is more expansionary in the flexible exchange rate case than in the fixed.

Assume now that capital is mobile. With the money supply held constant, fiscal expansion causes the interest rate to rise and capital to flow in from abroad. This in turn causes the exchange rate to appreciate (or depreciate less). For sufficiently mobile capital the appreciation will offset the expansionary effects of the initial fiscal intervention, i.e. fiscal policy is rendered nugatory by perfectly mobile international capital. On the other hand, if the exchange rate is fixed, this same internationally mobile capital will finance a trade deficit. Thus, fiscal policy is most effective at least in the short term with a fixed exchange rate and highly sensitive international capital flows.

See Also

- ▶ Deficit Financing
- ▶ Deficit Spending

Bibliography

The literature on budgetary policy is enormous. The following works contain useful bibliographies as well as being valuable in their own right.

Hansen, B. 1956. *The economic theory of fiscal policy*. London: Allen & Unwin.

Peacock, A.T. and Shaw, R. 1982. *The economic theory of fiscal policy*. 2nd edn, London: Allen & Unwin.

Peston, M. 1982. *Theory of macroeconomic policy*. 2nd edn, Oxford: Philip Allan.

Buffer Stocks

Ravi Kanbur

Sharp fluctuations in the prices of primary commodities seem to have been an integral part of the international economy for a long time. Keynes (1942) commented that 'One of the greatest evils in international trade before the war was the wide and rapid fluctuations in the world prices of primary products. . .' and went on to argue that a stable post-war economic order would require a scheme to stabilize commodity prices. In this context he suggested the setting up of buffer stocks

from which supply could be enhanced in periods of upward pressure on prices, and into which part of world supply could be withdrawn in periods of downward pressure on prices. As is documented in Volume XXVII of his collected works, Keynes lost the political battle to introduce such a buffer stock scheme. The post-war period saw attempts at price stabilization for individual commodities such as rubber, tin and sugar, but the major attempt at a comprehensive world wide scheme covering several commodities has been UNCTAD's 1976 proposal on an Integrated Program for Commodities. This proposal has also met with notable failure, and as things stand the case for buffer stocks on an international level seems to have lost its momentum. However, history teaches us that interest in buffer stocks is itself a cyclical phenomenon, and that concern for commodity price stabilization, particularly in the consuming nations of the North, is typically renewed in periods of commodity price booms.

To see the analytical arguments in favour of and against buffer stocks, consider the case where the demand curve for a commodity is fixed while the supply fluctuates for climatic or other reasons. Then the price of the commodity, determined by market forces, will also fluctuate. These fluctuations impose a cost on producers, but it is important to realize that (a) the root cause is supply uncertainty, and that (b) producers are interested not so much in price variability as in income variability. If the elasticity of demand is less than unity, then price fluctuations compensate for quantity fluctuations. In this situation, stabilizing price may actually increase income fluctuations. However, in the case where supply is certain but demand is variable, it is certain that price stabilization will also stabilize income. Newbery and Stiglitz (1981) carry out a comprehensive analysis of the benefits from price stabilization. Their conclusions are not supportive of buffer stock schemes: 'The major result of our analysis is to question seriously the desirability of price stabilization schemes, both from the point of view of the producer and of the consumer'.

So far as producers are concerned, the Newbery–Stiglitz conclusion follows from the observation that price changes typically compensate for supply fluctuations, and also because for observed values of the resultant income variability, and using experimental evidence on individual's risk aversion, the net 'risk premium' is not very large. The use of individual values for risk aversion, in evaluating the benefits to a nation of income stability, is questioned in Kanbur (1984). Also questioned is the direct identification of commodity earnings instability with instability of national income. Commodity earnings are foreign exchange, and in foreign exchange constrained regimes the cost of such instability may be much higher than viewing the instability in exactly analogous manner to an individual's insurance problem.

On the side of consumers, the Newbery–Stiglitz approach is again a microeconomic one, whereas the approach of Keynes (1942), Kaldor (1976) and Kanbur–Vines (1985) is macroeconomic in nature. Newbery and Stiglitz view the costs of commodity price instability to the consuming nations just like the cost to an individual of price instability – which turns out to be small. But the Keynesian argument centres around the role of commodity price instability in fuelling the inflationary spiral in the consuming nations. Kanbur and Vines (1985) consider the benefits of stabilization in such a Keynesian setting, and find them to be larger than hitherto supposed.

The debate on buffer stocks will no doubt continue, but there are some lessons which have certainly been learned from the latest round of analysis. Firstly, supply responses are important. It is crucial to an understanding of the operation of buffer stocks to realize that a change in the degree of price instability will change supply conditions and hence the environment in which the scheme has to operate. This simultaneity must be taken into account. Secondly, the behaviour of agents other than producers has to be analysed – in particular, buffer stock schemes may be open to speculative attack rather in the manner of a central bank attempting to maintain a fixed exchange rate with limited reserves. Thirdly, the optimal behaviour of the buffer stock authority is a complicated matter. One approach to characterizing the solution, for a given economic environment, is to use stochastic dynamic programming; the analysis is further

complicated by the fact that the solution will itself affect supply response and hence the economic environment. Also, the solution may be difficult to implement operationally. In actual operations, a 'band width rule' is often specified, which gives authority to the buffer stock managers to intervene above and below critical price levels.

Finally, there remains the question of why international negotiations on a comprehensive system of buffer stocks for major commodities have been singularly unsuccessful. Does this indicate that the benefits from commodity price stabilization are small? We have to be careful in separating out the intellectual pros and cons of stabilization from the political realities of international negotiations. For a start, in these negotiations the *stability* of prices gets confounded with the *level* of prices. The latter is a matter of the distribution of income between producers and consumers, while it is the former which has been the focus of most analysis. Also, while attempts to tie together a number of commodities in a single scheme can be justified from the point of view of being able to take advantage of the covariance between different prices, such tying together inevitably raises conflicts within the producing nations and within the consuming nations. Finally, discussions of and negotiations on UNCTAD's Integrated Program for Commodities inevitably became entangled with wider questions on the New International Economic Order. For these reasons it would be inappropriate to read the failures of such proposals to find acceptance as a sure sign of an intellectual weakness in the case for buffer stocks. The debate will continue, and the issue is still wide open.

See Also

- ▶ Commodity Reserve Currency
- ▶ Inventories

References

Kaldor, N. 1976. Inflation and recession in the world economy. *Economic Journal* 86: 703–714.

Kanbur, S.M.R. 1984. How to analyse commodity price stabilization? A review article. *Oxford Economic Papers* 36(3): 336–358.

Kanbur, S.M.R., and D.A. Vines. 1987. North–South interaction and commod control. *Journal of Development Economics*.

Keynes, J.M. 1942. The international regulation of primary products. Reprinted in *Collected Writings of John Maynard Keynes*, vol. XXVII. London: Macmillan, 1982.

Newbery, D.M.G., and J.E. Stiglitz. 1981. *The theory of commodity price stabilization – A study in the economics of risk*. Oxford: Oxford University Press.

Built-In Stabilizers

Joseph A. Pechman

Built-in stabilizers are automatic fiscal adjustments that reduce the national income multiplier and thus cushion the effect of changes in autonomous spending on the level of income. Suppose the multiplier is $1/(1 - c)$ in an economy with no tax or with a lump sum tax, where c is the marginal propensity to consume. With a proportional income tax, t, the multiplier is reduced to $1/_c(1–c(1–t)]$.

The two groups of stabilizers are taxes, in particular income taxes, and government transfer payments, such as unemployment compensation and welfare benefits. These stabilizers moderate the fall in income when private spending declines and restrain the increase in income when private spending rises. The properties of built-in stabilizers were discovered by many people soon after John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936) was published, but the first person to use the term was Albert Gailord Hart (in De Chazeau et al. 1946); Hart (1945), Brown (1955), Richard A. Musgrave (1948) and Herbert Stein (De Chazeau et al. 1946) played a major role in the development and popularization of the theory of built-in stabilizers.

Two related statistical measures are used to describe fiscal responses to changes in economic activity. *Built-in flexibility* (b) is the change in tax revenues per unit of change in income dT/

d*Y*. *Elasticity* (*e*) is the ratio of the percentage change in tax revenues to the percentage change in income $[dT/T \div dY/Y] = (dT/dY)(Y/T)$. Denoting the effective rate *T/Y* as *r*, *b* can be expressed as a function of *e*: $b = e \times r$. Thus, the effectiveness of a tax (or expenditure) programme in cushioning changes in income is greater the higher its elasticity and the higher its effective rate. During periods of inflation, the increase in tax revenues or reduction in expenditures must be in *real* terms for the tax or expenditure programmes to qualify as automatic stabilizers.

Because of built-in stabilizers, the actual deficit or surplus reflects the prevailing levels of income and unemployment, as well as the government's fiscal policy. Thus, the effects of various fiscal programmes on demand may be compared only after removing the effects of the built-in stabilizers on the budget. By convention, the calculation is made at a high level of employment (say, 94 or 96% of the labour force, depending on which rate is consistent with non-accelerating inflation) and the result is called the *high-employment deficit or surplus*. The stabilizing budget policy of the Committee for Economic Development, a non-profit organization of influential businessmen and educators, proposed that tax rates should be set to balance the budget or yield a small surplus at high employment.

The individual income tax is the most important stabilizer, both because of its size and progressive rate structure. When incomes fall, some people who were formerly taxable drop below the taxable level; others are pushed into lower tax brackets. When incomes rise, more people become taxable and others move into higher tax brackets. The result is that the yield of the individual income tax rises and falls more than in proportion to changes in income. Since consumption depends to a considerable extent on disposable personal income, automatic changes in individual income tax liabilities keep consumption more stable than it otherwise would be.

The cyclical elasticity of the corporation income tax is greater than that of the individual income tax because corporate profits fluctuate more widely than individual incomes. However, the corporate tax is less important as a stabilizer because it is much smaller. The policy of corporations to cut into saving rather than to reduce dividends when profits decline also stabilizes final demand. The reduction of retained corporate earnings prevents a corresponding decline in disposable personal income, thus helping to maintain spending of consumers.

Receipts from a general consumption tax (such as a general sales tax) or a proportional payroll tax respond about in proportion to changes in income. Excise taxes are even less effective automatic stabilizers than a general consumption tax because they are usually levied on the number of units (for example, cents per gallon) rather than as a percentage of the value of purchases and thus do not increase as prices increase.

The major built-in stabilizer on the expenditure side of the budget is unemployment compensation. These payments maintain consumption as output and employment fall. As incomes go up and employment increases, unemployment compensation declines. Other transfer payments (for example, social security and welfare benefits) also increase more rapidly during recessions as the number of beneficiaries increases, but their cyclical fluctuations are much smaller than the fluctuations of unemployment benefits.

Built-in flexibility is clearly desirable when inflation results from an overheated economy. But it may have perverse effects if prices are rising when employment and output are falling. In these circumstances, real tax receipts will rise if nominal income elasticity exceeds unity. Such a tax increase would aggravate the decline in economic activity.

Automatic increases in income taxes during periods of inflation raise real tax burdens without any action on the part of legislatures. To offset this 'bracket creep', many countries adjust income tax rates and exemptions automatically for inflation. In these indexed systems, the stabilizing effect of the income tax is limited to the change in tax liabilities from changes in real income, and not from the inflation component of income changes. Less attention has been paid in recent years to built-in flexibility, primarily because of the opposition to automatic increases in real tax burdens as a result of inflation.

See Also

► Stabilization Policy

References

Brown, E.C. 1955. The static theory of automatic fiscal stabilization. *Journal of Political Economy* 63: 427–440.

Committee for Economic Development. 1947. *Taxes and the budget: A program for prosperity in a free economy.* New York: Committee for Economic Development.

De Chazeau, M.G., et al. 1946. *Jobs and markets.* A Committee for Economic Development Research Study. New York: McGraw-Hill.

Hart, A.G. 1945. Model building and fiscal policy. *American Economic Review* 35: 531–558.

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

Musgrave, R.A., and M.H. Miller. 1948. Built-in flexibility. *American Economic Review* 38: 122–128.

Slitor, R.E. 1948. The measurement of progressivity and built-in flexibility. *Quarterly Journal of Economics* 62: 309–313.

Bukharin, Nikolai Ivanovitch (1888–1938)

Donald J. Harris

Keywords

Austrian economics; Bukharin, N. I.; Capitalism; Falling rate of profit; Imperialism; Lenin, V. I.; Luxemburg, R.; Marginal revolution; Marx, K. H.; Monopoly; Period of production; Planning; Socialism; Underdevelopment; Uneven development

JEL Classifications

B31

Nikolai Bukharin is commonly acknowledged to have been one of the most brilliant theoreticians in the Bolshevik movement and an outstanding figure in the history of Marxism. Born in Russia, he studied economics at Moscow University and (during 4 years of exile in Europe and America) at the Universities of Vienna and Lausanne (Switzerland), in Sweden and Norway and in the New York Public Library. While still a student, he joined the Bolshevik movement. Upon returning to Russia in April 1917, he worked closely with Lenin and participated in planning and carrying out the October Revolution. After the victory of the Bolsheviks he proceeded to assume many high offices in the Party (becoming a member of the Politbureau in 1919) and in other important organizations. In these various capacities he came to exercise great influence within both the Party and the Comintern. Under Stalin's regime, however, he lost most of his important positions. Eventually, he was among those who were arrested and brought to trial under charges of treason and was executed on 15 March 1938.

At the peak of his career Bukharin was regarded as the foremost authority on Marxism in the Party. He was a prolific writer: there are more than five hundred items of published work in his name, most of them written in the hectic 12-year period 1916–1928 (for a comprehensive bibliography, see Heitman 1969). Only a few of these works have been translated into English and these are the works for which he is now most widely known. A brief description of the major items gives an indication of the scope and range of his intellectual interests.

The Economic Theory of the Leisure Class (1917) is a detailed and comprehensive critique of the ideas of the Austrian school of economic theory, as represented by the work of its chief spokesman Eugen von Böhm-Bawerk, but situated in the broader context of marginal theory as it had appeared up to that time. In *Imperialism and World Economy* (1918) he formulated a revision of Marx's theory of capitalist development and set out his own theory of imperialism as an advanced stage of capitalism. This was written in 1914–15, a year before Lenin's *Imperialism*, and is credited with having been a major influence on Lenin's formulation. The theoretical structure of the argument is further elaborated in *Imperialism and the Accumulation of Capital* (1924) by way of a critique of the ideas of Rosa Luxemburg, another leading Marxist writer of that time. *The ABC of*

Communism (1919), written jointly with Evgenii Preobrazhensky and used as a standard textbook in the 1920s, is a comprehensive restatement of the principles of Marxism as applied to analysis of the development of capitalism, the conditions for revolution, and the nature of the tasks of building socialism in the specific context of the Soviet experience. This book, taken with his *Economics of the Transition Period* (1920), constitutes a contribution to both the Marxist theory of capitalist breakdown and world revolution on the one hand and the theory of socialist construction on the other. *Historical Materialism*: *A System of Sociology* (1921), another popular textbook, combines a special interpretation of the philosophical basis of Marxism with what is perhaps the first systematic theoretical statement of Marxism as a system of sociological analysis. In style much of this work is highly polemical and geared to immediate political goals. But it reveals also a versatility of intellect, serious theoretical concern, and scholarly inclination. Arguably, his works represent in their entirety 'a comprehensive reformulation of the classical Marxian theory of proletarian revolution' (Heitman 1962, p. 79). Viewed from the standpoint of their significance in terms of economic analysis, three major components stand out.

There is, first, the critique of 'bourgeois economic theory' in its Austrian version. Bukharin's approach follows that which Marx had adopted in *Theories of Surplus Value,* which is to give an 'exhaustive criticism' not only of the methodology and internal logic of the theory but also of the sociological and class basis which it reflects. He scores familiar points against particular elements of the theory, for instance, that utility is not measurable, that Böhm-Bawerk's concept of an 'average period of production' is 'nonsensical', that the theory is static. Such criticisms of the technical apparatus of the theory have since been developed in more refined and sophisticated form (see Harris 1978, 1981; Dobb 1969). Moreover, certain weaknesses in Bukharin's presentation, such as an apparent confusion between marginal and total utility and misconception of the meaning of interdependent markets, can now be readily recognized. But these are matters that were not well understood at the time, even by exponents of the theory. Bukharin views them as matters of lesser importance. What is crucial for him is 'the point of departure of the ... theory, its ignoring the social-historical character of economic phenomena' (1917, p. 73). This criticism is applied with particular force to the treatment of the problem of capital, the nature of consumer demand, and the process of economic evolution. As to the sociological criticism, his central thesis is that the theory is the ideological expression of the rentier class eliminated from the process of production and interested solely in disposing of their income through consumption. This thesis can be faulted for giving too mechanical and simplistic an interpretation of the relation between economic theory and ideology where a dialectical interpretation is called for (compare, for instance, Dobb 1973, ch. 1, and Meek 1967). But the issue of the social-ideological roots of the marginal revolution remains a problematic one, as yet unresolved, with direct relevance to current interest in the nature of scientific revolutions in the social sciences (see Kuhn 1970; Latsis 1976).

Secondly, Bukharin's work clearly articulates a conception of the development of capitalism as a world system to a more advanced stage than that of industrial capitalism which Marx had earlier analysed. This new stage is characterized by the rise of monopoly or 'state trusts' within advanced capitalist states, intensified international competition among different national monopolies leading to a quest for economic, political and military control over 'spheres of influence', and breaking out into destructive wars between states. These conditions are seen as inevitable results deriving from inherent tendencies in the capitalist accumulation process, at the heart of which is a supposed falling tendency in the overall average rate of profit. Altogether they are viewed as an expression of the anarchic and contradictory character of capitalism. The formation of monopolies is supposed to take place through reorganization of production by finance capitalists as a way of finding new sources of profitable investment and of exercising centralized regulation and control of

the national economy. This transformation succeeds for a time at the national level but only to raise the contradictions to the level of the world economy where they can be resolved only through revolutions breaking out at different 'weak links' of the world-capitalist system. The idea of a necessary long-term decline in the rate of profit, and also the specific role assigned to financial enterprises as such, can be disputed. A crucial ingredient of the argument is the idea of oligopolistic rivalry and international mobility of capital as essential factors governing international relations. In this respect the argument anticipates ideas that are only now being recognized and absorbed into the orthodox theory of international trade and which, in his own time, were conspicuously neglected within the entire corpus of existing economic theory. Much of the analysis as regards a necessary tendency to uneven development between an advanced *centre* and underdeveloped *periphery* of the world economy has also been absorbed into contemporary theories of underdevelopment. Underpinning the whole argument is a curious theory of 'social equilibrium' and of 'crisis' originating from a loss of equilibrium. 'To find the law of this equilibrium', he suggests (1920, p. 149), 'is the basic problem of theoretical economics and theoretical economics as a scientific system is the result of an examination of the entire capitalist system in its state of equilibrium'.

The third component is a comprehensive conception of the process of socialist construction in a backward country. These ideas came out of the practical concerns and rich intellectual ferment associated with the early period of Soviet development but have a generality and relevance extending down to current debates both in the development literature and on problems of socialist planning. The overall framework is one that conceives of socialist development as a long-drawn-out process 'embracing a whole enormous epoch' and going through four revolutionary phases: ideological, political, economic and technical. The process is seen as occurring in the context of a kind of war economy involving highly centralized state control, though there is an optimistic prediction of an ultimate 'dying off of the state power'. Room is allowed for preserving and maintaining small-scale private enterprise. The agricultural sector is seen as posing special problems, due to the assumed character of peasant production, which can only be overcome through transformation by stages to collectivized large-scale production. Even so, it is firmly held (in 1919) that 'for a long time to come small-scale peasant farming will be the predominant form of Russian agriculture', a view which Bukharin later abandoned in support of Stalin's collectivization drive. In industry, too, small-scale industry, handicraft, and home industry are to be supported, so that the all-round strategy is one that seems quite similar to that of 'walking on two legs' later propounded by Mao for China. An extensive discussion is presented of almost every detail of the economic programme, from technology to public health, but little or no attention is given to issues of incentives and organizational problems of centralization/decentralization which have emerged as crucial considerations in later work.

Cohen (1973) remains a classic biography; his widow's memoirs, Larina (1993) are also of interest.

Selected Works

1917. *The economic theory of the leisure class.* New York: Monthly Review Press, 1972.
1918. *Imperialism and world economy.* New York: Monthly Review Press, 1973.
1919. (With E. Preobrazhensky.) *The ABC of communism.* Harmondsworth: Penguin Books, 1969.
1920. The economics of the transition period. In *The politics and economics of the transition period,* ed. K.J. Tarbuck. London: Routledge & Kegan Paul, 1979.
1921. *Historical materialism, a system of sociology.* Ann Arbor: University of Michigan Press, 1969.
1924. *Imperialism and the accumulation of capital.* New York: Monthly Review Press, 1972.

Bibliography

Cohen, S. 1973. *Bukharin and the Bolshevik revolution: A political biography: 1888–1938*. New York: Random House.
Dobb, M. 1969. *Welfare economics and the economics of socialism*. Cambridge: Cambridge University Press.
Dobb, M. 1973. *Theories of value and distribution since Adam Smith*. Cambridge: Cambridge University Press.
Harris, D.J. 1978. *Capital accumulation and income distribution*. Stanford: Stanford University Press.
Harris, D.J. 1981. Profits, productivity, and thrift: The neoclassical theory of capital and distribution revisited. *Journal of Post-Keynesian Economics* 3 (3): 359–382.
Heitman, S. 1962. Between Lenin and Stalin: Nikolai Bukharin. In *Revisionism*, ed. Leopold Labedz. New York: Praeger.
Heitman, S. 1969. *Nikolai I. Bukharin: A bibliography.* Stanford: Hoover Institution.
Kuhn, T. 1970. *The structure of scientific revolutions.* 2nd edn, enlarged. Chicago: University of Chicago Press.
Larina, A. 1993. *This I cannot forget: The memoirs of Nikolai Bukharin's widow. New York*: *W. W. Norton*.
Latsis, S., ed. 1976. *Method and appraisal in economics*. Cambridge: Cambridge University Press.
Lenin, V.I. 1917. *Imperialism, the highest stage of capitalism*, 1939. New York: International Publishers.
Meek, R.L. 1967. *Economics and ideology and other essays*. London: Chapman & Hall.

Bullionist Controversies (Empirical Evidence)

Lawrence H. Officer

Abstract

There are three historical episodes in which bullionist–anti-bullionist macroeconomic debates occurred: Sweden (1745–76), England (1797–1821), and Ireland (1797–1821). Expressing the bullionist and anti-bullionist models as chains of causation facilitates presentation of empirical studies of the experiences. As a group, the studies suggest that the anti-bullionist position is more supported by the empirical evidence.

Keywords

Bank of England; Bank or Ireland; Bank Restriction Period; Bullionist controversies; Cointegration; Continental System; Copper standard; Fixed exchange rates; Floating exchange rates; Gold standard; Granger causality; Monetarism; Monetary base; Purchasing power parity; Quantity theory of money; Real bills doctrine; Riksbank; Silver standard; Spurious regressions; Time series analysis

JEL Classifications

N2

The bullionist periods of Sweden, England, and Ireland involved bullionist–anti-bullionist macroeconomic debates, with empirical studies vindicating largely the anti-bullionist side.

History of Bullionist Periods

The bullionist controversy is a debate that can occur in monetary history when a paper currency and floating exchange rate interrupt a metallic standard. The three famous bullionist periods pertain to Sweden, England and Ireland. In 1745, the Riksbank made its notes inconvertible into copper bullion, resulting in the paper daler. It was not until 1776 that the Swedish bullionist period ended, with conversion to a new currency unit (the riksdaler) on a silver standard. The English, followed by the Irish, bullionist period began in 1797, each by government order requiring the Bank of England and Bank of Ireland to cease making gold payments for its notes. Legislation, periodically renewed, solidified the orders. In 1821 the Bank of England, followed by the Bank of Ireland, resumed payment in gold, and the countries were back on a gold standard. The English episode is called the 'Bank Restriction Period'.

The three bullionist periods involved common elements: a prior metallic standard replaced by a paper standard, a fixed exchange rate (constrained within a band around an effective mint parity)

giving way to a floating rate, unusually high inflation, depreciation of the currency in the foreign-exchange and bullion markets, a sub-period of deflation, and eventual return to a specie standard and fixed exchange rate. Also, periods of war occurred both before and during the bullionist periods.

Some characteristics were shared by only two of the periods. First, the proximate cause of the Swedish and English Restrictions was a tremendous loss of reserves on the part of the Riksbank and Bank of England. This was not the case for the Bank of Ireland; British pressure induced the Irish government to suspend convertibility of Bank of Ireland notes. Second, for Sweden and England, their main trading partners remained on a metallic standard. This was not so for Ireland, with England also on paper. Third, England and Ireland returned to a gold standard at the old parity; Sweden switched from an effective copper to an effective silver standard, and banknotes were depreciated by 50 per cent in terms of silver.

Two additional features characterize all three periods. First, the macroeconomic debate centred on determination of the exchange rate and price level, and their relationship to the balance of payments and note issues of the central bank. The bullionists adopted a monetarist approach, and the anti-bullionists a non-monetarist position. Second, Parliament played a key role in the controversy. In the case of Sweden, two political parties vied for control of Parliament. The 'Caps' had a bullionist agenda, and the 'Hats' an anti-bullionist policy. Both had intellectual supporters on the outside. The British House of Commons appointed committees, in 1804 and 1810, to investigate the depreciated Irish and English currencies. Each committee produced a highly bullionist report, important in the literature; but in neither case was the report favourably received by Parliament.

Bullionist, Anti-bullionist, and Country-Bank Models

To examine the empirical literature on the bullionist controversies, each side is represented by its mainstream model of chains of causality, sequential hypotheses. Notation is $X \rightarrow Y$ ('X causes Y, with $\partial Y/\partial X > 0$'). Multiple hypotheses are W, $X \rightarrow Y$ ('$W \rightarrow Y$ and $X \rightarrow Y$') and $X \rightarrow Y, Z$ ('$X \rightarrow Y$ and $X \rightarrow Z$'). The subscript f designates a foreign variable. Variables are:

BN: central-bank notes in circulation
BP: balance-of-payments deficit
CN: country banknotes in circulation
ER: exchange rate, price of foreign currency
FR: remittances to foreign countries
HQ: quantity and quality of harvest
MS: money supply (M1)
PG: price of gold
PL: price level
PM: price of imports
PW: price of wheat
TR: foreign trade restrictions

The *bullionist model* is decidedly monetarist: only monetary variables affect only monetary variables. The English-bullionist chain of causation is:

$$BN \rightarrow MS \rightarrow PL \rightarrow ER, PG.$$

$BN \rightarrow MS$ reflects the bullionist, and correct, perception that Bank of England notes constituted the monetary base during the Restriction Period. There was a hierarchy of banks: the Bank of England (central bank), London private banks, and country banks. Bank of England notes (held as reserves by the country banks and London private banks) were non-redeemable; deposits at the Bank (held as reserves only by the London private banks) were cashable only in Bank of England notes. The country banks – but not the London private banks – issued notes. There were no legal reserve requirements for any bank; but, like all companies, banks had to settle their debts (note and deposit liabilities) in cash. Reserves of the country banks were principally deposits at the London private banks, with Bank of England notes (and, in principle, gold) for vault cash. Bank of England notes circulated in and around London, as well as in Lancashire and Norwich; country banknotes circulated elsewhere in

England and Wales. During the Bank Restriction Period, the English country banks and Scottish banks 'redeemed' their notes in Bank of England notes rather than gold. This was a matter of practice rather than law.

Strictly speaking, gold coin was a component of the monetary base, but the premium on gold bullion did not have a counterpart in the premium of gold coin over Bank of England notes. There was no legal market for domestic coin in terms of paper money, and an overwhelming proportion of the gold coin nominally in circulation or newly minted was in fact hoarded or exported.

For the bullionists (and anti-bullionists), the money supply had as components Bank of England notes, country banknotes, and coin. In excluding deposits from M1, the writers of the Restriction Period were not far off the mark. First, except in London, 'deposits' generally meant time or savings deposits rather than demand deposits. Second, if interbank transactions are excluded, demand deposits typically were exchanged for cash rather than transferred to another account.

BN → MS was also asserted by the Irish bullionists, even though the banking system was looser. In and around Dublin, notes of the Dublin private banks circulated along with notes of the Bank of Ireland. Gold did not circulate, except in the north until 1808–9, when it was replaced by the notes of newly established Belfast banks. Elsewhere, local private banknotes generally dominated, but in competition with Bank of Ireland notes and, to a lesser extent, Dublin private-bankers' notes. The private banks kept their reserves in Bank of Ireland notes (and gold), and by convention their notes were redeemed in Bank of Ireland notes.

In the Swedish bullionist period, BN = MS. With little coin circulating, no commercial banks in existence, and deposits at the Riksbank representing merely the right to make withdrawals in notes, Riksbank notes essentially equalled the money supply.

MS → PL pertains to the quantity theory of money. Underlying this theory is the bullionist view that the Bank of England effectively pegged the market interest rate at five per cent, by standing ready to discount all 'good' commercial bills at that rate. Thus the monetary base is perfectly elastic at the constant discount rate of five per cent, a powerful impetus to the quantity theory.

There is good reason for this view: the usury laws set a five per cent limit on annual interest on bills of exchange, and the discount rate of the Bank of England was fixed at this rate. While bill brokers could charge a commission and private banks could require a minimum balance, the Bank did not use such devices. The market discount rate (for good bills) did not exceed five per cent during the Restriction. In fact, only for about a year (beginning July 1817) did the market rate even fall below five per cent. The situation was yet stronger regarding the Bank of Ireland. Its discount rate was limited to five per cent by charter.

However, the English and Irish bullionists were wrong in inferring that the monetary base (essentially BN) could rise without limit. First, there is evidence that in historical fact the monetary base was not perfectly elastic. Only 'good' bills—a minority of bills—were acceptable by the Banks. Also, the Bank of England effectively regulated discounts via a rationing system. These facts act against the quantity theory but support the concept of BN as an autonomous policy variable.

Second, even if the *supply* of the monetary base (essentially BN) is perfectly elastic at the pegged market interest rate, BN is limited by the *demand* for the monetary base. The Bank of England and Bank of Ireland could not induce the private sector to hold more BN than demanded. BN was viewed by the bullionists as the first link in the causal chain; but it is an endogenous variable. A low level of economic activity could hold down the demand for BN.

PL → ER is the purchasing-power-parity theory (given PLf), the causal nature of which is generally ignored in the modern literature. PL → PG involves a relatively unchanged PGf, for, under perfect markets, PG is the product of ER and PGf. PG was not as interesting to the Swedish and Irish bullionists as it was to the English. Sweden had been on a copper standard; the concern in Ireland was depreciation of the Irish

currency against the British. For the Swedish and English protagonists, foreign exchange was Continental currencies.

For most Swedish and Irish bullionists, the latter part of the chain is merely MS → PL , E-R. The price level and exchange rate are co-determined by the money stock. Some Irish bullionists allowed for a changing foreign (English) price level, so the hypothesis becomes MS/MS_f (or BN/BN_f) → ER.

The English *anti-bullionist model* involves a balance-of-payments theory of the exchange rate, with demand for and supply of bills of exchange represented by the payments deficit (BP), yielding ER and PG. The state of the harvest, a real factor, determines the domestic price of grain, represented by the price of wheat (PW). The exchange rate is an ingredient in the price of imports, which, together with PW, determines PL. These anti-bullionists saw three principal determinants of BP, that is, of shifts in the demand for or supply of foreign exchange: PW, foreign trade restrictions (wartime restraints: the Continental System and the American embargo), and foreign remittances (external government payments: direct military expenditure and subsidies to allied countries). The English anti-bullionist causal chain is:

$$\begin{array}{ccccc} 1/HQ & \rightarrow PW & \rightarrow & PL \rightarrow B/N \\ & \downarrow & & \uparrow \\ TR, FR & \rightarrow BP \rightarrow & ER, PG & \rightarrow PM \end{array}$$

In emphasizing the price of wheat, the anti-bullionists recognized the highly agrarian state of the British economy, notwithstanding the industrial revolution in progress. The emphasis on wartime interference with trade and on external military expenditure reflected the French Revolutionary and Napoleonic Wars, in which Britain was engaged for much of the Bank Restriction Period.

For the Irish anti-bullionists, concerned with the English exchange, TR and PG were unimportant. They did not make explicit the connection of PW and PM to PL, and FR took the form of payments to absentee landlords in England. Some consolidated the trade balance, interest payments, net capital exports, and FR, to compose (and presumably shift) BP in the causal chain. They left unclear the mechanism from BP to PL. The Swedish anti-bullionists had the chain: BP → ER → PM → PL, allowing real shocks to operate on BP.

The anti-bullionists used the 'real-bills' doctrine to reverse the bullionist BN → PL causation. They accepted that the Bank behaved passively in its note issuance, but used the real-bills theory to demonstrate that excess issue (beyond the 'needs of trade') would be returned to the Bank instead of acting to increase the price level monetarily. Only non-monetary forces could cause real income and then the price level to increase, and would underlie the demand for discounting to finance a higher volume of transactions, whence PL → BN. The Irish bullionists also propounded the real-bills doctrine (for the Bank of Ireland), although some saw ER playing the role of PL.

Bullionists in all three periods essentially inverted the real-bills theory by offering the *policy rule* that central-bank note issuance should be oriented to the exchange rate and (for the English bullionists) gold price: ER, PG → 1/BN.

Extension to Country Banks

A subsidiary part of the English and Irish bullionist controversies was the extent to which the country banks (in Ireland, including Dublin private banks) could affect the money supply independent of the central bank. Should the first hypothesis in the bullionist chain, BN → MS, incorporate CN naturally as BN → CN → MS (country banks unable to vary their note issues independent of the central bank)? Or should the hypothesis be (BN + CN) → MS (the central bank and country banks able either jointly or separately to change their issues)? Or should the hypothesis be CN → MS (only the country banks, not the central bank, having the power to change the money supply)? The question was answered differently by groups that cut across the bullionist–anti-bullionist line.

The correct hypothesis is not clear, because of the environment in which banks operated. Among

the complicating, and largely unknown, elements are the extents to which (*a*) one-time replacement of gold by central-bank notes in reserves altered country-bank policy regarding reserve ratios, (*b*) country-bank reserve ratios varied over time, (*c*) public preference for central-bank over country-bank notes changed in particular geographic areas and over time, (*d*) circulation of counterfeit notes and unlicensed-bank notes affected the demand for and supply of country-bank and central-bank notes, and (*e*) London private banks were prepared to run down their reserve ratios to accommodate country-bank demand for additional reserves.

Empirical Studies: Visual Comparison of Movements of Variables

The empirical studies examined here make use of quantitative information to test one or more component hypotheses of the bullionist or anti-bullionist models. It is logical to begin with contemporary studies, as it is the hypotheses of contemporary authors that are delineated in the previous sections.

All contemporary investigations use a simple technique: visual inspection of sets of figures, formal tables, or charts. The earliest such studies pertain to the Ireland bullionist period, with BN and BN_f the note circulations of the Bank of Ireland and Bank of England. Parnell (1804), Foster (1804) and the 1804 Currency Report (in Fetter 1955) find that BN → ER is confirmed. Ó Gráda (1993) and Fetter (1955) criticize the Report for its small number of observations and selective observations. These criticisms can be extended to Parnell, but not to Foster. The report of 1804 and Parnell also claim successful testing of BN/BN_f → ER. Ó Gráda (1991) finds this part of the Report misleading in several respects; but the Report is to be commended for making specific allowance for the replacement of gold coin by notes. The Report also claims to disprove BP → ER, via computation of a net balance-of-payments surplus. However, this proves little, because there is no representation of shifts in the demand for or supply of bills on London.

Contemporary empirical work on the English bullionist period begins with Ricardo (1811), whose positive finding of BN → ER (Hamburg exchange) is reinforced by observation of a lagged effect and by accounting for replacement of gold coin by Bank of England notes. Galton (1813) confirms that BN → ER, PG. Anonymous (1819) sees mixed evidence for that hypothesis, but observes that grain imports and FR (not precisely defined) affect the exchange rate – the first results in favour of anti-bullionism.

There is a hiatus of more than a century, but three groupings of subsequent work do not merit review. First is any investigation, such as Silberling (1924), involving the London price of the Spanish dollar to represent the exchange rate. That choice is methodologically unsound. Britain was on a suspended gold (not silver) standard, and the Spanish silver dollar was not a circulating coin in Hamburg, the main foreign-exchange market. Second are tests making use of Silberling-developed series of Bank of England total advances and their private versus public components. These series have been shown to be seriously inconsistent with the Bank's published data. Third, and most unfortunate, are all studies using 'data' on country banknote circulation. There exist no true data on country banknote circulation in England, or private banknote circulation in Ireland, during the bullionist period. Further, with no legal or fixed reserve ratio of note liabilities to cash, the circulation of the Bank of England, or Bank of Ireland, cannot be used to infer that of the private banks.

Private banks were required to register at the Stamp Office and pay a stamp tax on notes prior to issuance. Some have used stamp-tax data to develop proxy CN series for England, based on the value of country banknotes stamped; but the series are based on assumptions so tenuous as to make the series unusable.

Silberling (1924) develops an annual series for FR ('extraordinary foreign payments'), consisting of grain imports over a normal amount, Continental British war expenditures, and subsidies to foreign states. Using various definitions of FR, based largely on Silberling, Angell (1926) shows that FR → ER, but can find no causal relationship between PL and ER. This result, favourable to

anti-bullionism, is supported by Morgan (1939, 1943) and Viner (1937). Morgan rejects BN → PL, but accepts PL → BN. His only finding not supportive of anti-bullionism is the lack of a relationship between PW and PL or BN.

Gayer et al.(1953, p. 932) support BP → ER; but they represent BP by the balance of trade, the data of which are crude. For the Swedish period, Eagly (1971) and Bernholz (1982, 2003) support BN → PL, ER, favourable to bullionism.

This entire body of literature must be viewed with caution. First, interpretation of relationships among variables is subjective when data are merely tabulated or plotted. Second, macroeconomic variables are generally non-stationary, leading to the possible outcome of 'spurious regression'.

Empirical Studies: Time-Series Analysis

Myhrman (1976) computes annual growth rates of BN and PL, for Sweden and England, and argues that BN → PL. Jonung (1976) does the same for Sweden alone. Transforming data to growth rates could yield stationarity. In a joint test of bullionist and anti-bullionist hypotheses, Arnon (1990) regresses PL on PW, BN, and a trend. He finds that BN contributes more to the regression than PW. The variables are transformed to correct for serial correlation, which could correct spurious regression.

Formal time-series analysis in the bullionist literature begins with Ó Gráda (1989, 1993). For England, he cannot reject a cointegration relationship between logPL and logBN. This means that there is no long-term equilibrium between the variables, a failure of support for either bullionism or anti-bullionism The same negative result holds for Ireland, with BN/BN_f used in place of BN.

Nachane and Hatekar (1995) use Granger causality and cointegration techniques for England. Their variables are PL, ER, PG, BP, and BN/Y (transformed to logarithms except for BP, the only non-stationary variable), where Y is real output. Their results are ER → PL , PL → BN/Y (with PL and BN/Y the only cointegrated pair of variables), and BP → ER, PG. The findings are strongly supportive of anti-bullionism; but measuring the money supply in relation to output is outside the mainstream controversy.

The analyses of Ó Gráda and Nachane–Hatekar are restricted to bivariate econometrics. Officer (2000) applies multivariate testing to PL, ER, BN, FR, and PW, for England. Non-stationarity cannot be rejected, but cointegration is rejected. The logarithmic variables are first-differenced (to achieve stationarity), and Granger causality testing along with innovation analysis is applied. Results are mixed for bullionism, but unambiguously favourable to anti-bullionism. For example, the real-bills doctrine, PL → BN, receives stronger support than does the quantity theory, BN → PL.

It is logical that the time period for testing hypotheses be strictly within the pertinent bullionist period, because the alternative (bullionist versus anti-bullionist) models are geared to a paper standard and floating exchange rate. As his sample, Officer uses the 96 quarters encompassed by the Bank Restriction Period (1797–2 to 1821–1). Nachane and Hatekar employ annual data, and extend the time period to 1838. Ó Gráda has quarterly observations, but begins his time periods prior to 1797.

Nachane and Hatekar can also be criticized for using the exchange rate on Paris rather than Hamburg to represent ER. There are no quotations on Paris until 1802 (whence they lose observations), and historians agree that the Hamburg exchange was more representative during wartime.

To conclude: certainly, at least for England, the anti-bullionist position receives greater support (or less contradiction) than the bullionist side of the controversy. This result is inconsistent with modern macroeconomics. The anti-bullionist approach to the exchange rate (a flow theory) and monetary policy (passive, accommodating the price level) has been superseded in modern theory. Also, modern monetarism emanates from bullionism.

See Also

▶ Cointegration
▶ Granger–Sims Causality

- Monetarism
- Purchasing Power Parity
- Quantity Theory of Money
- Real Bills Doctrine
- Real Bills Doctrine Versus the Quantity Theory
- Spurious Regressions

Bibliography

Angell, J.W. 1926. *The theory of international prices*. Cambridge, MA: Harvard University Press.

Anonymous. 1819. Two tables...showing the rates of exchange on Hamburgh compared with the amount of bank notes, and the price of gold, and with the foreign expenditure, and the value of grain imported from the year 1793 to 1819. *Pamphleteer* 15: 281–286.

Arnon, A. 1990. What Thomas Tooke (and Ricardo) could have known had they constructed price indices. In *Keynes, macroeconomics and method*, ed. D.E. Moggeridge. Aldershot: Edward Elgar.

Bernholz, P. 1982. *Flexible exchange rates in historical perspective*. Princeton: International Finance Section, Princeton University.

Bernholz, P. 2003. *Monetary regimes and inflation*. Cheltenham: Edward Elgar.

Eagly, R.V. 1971. *The swedish bullionist controversy*. Philadelphia: American Philosophical Society.

Fetter, F.W. 1955. *The irish pound, 1797–1826*. Evanston: Northwestern University Press.

Foster, J.L. 1804. *An essay on the principles of commercial exchanges*. London: J. Hatchard.

Galton, S.T. 1813. *A chart, exhibiting the relation between the amount of Bank of England notes in circulation, the rate of foreign exchanges, and the price of gold and silver bullion and of wheat*. London: J. Johnson.

Gayer, A.D., W.W. Rostow, and A.J. Schwartz. 1953. *The growth and fluctuation of the british economy, 1790–1850*. Oxford: Clarendon Press.

Jonung, L. 1976. Money and prices in Sweden, 1732–1972. *Scandinavian Journal of Economics* 78: 40–58.

Morgan, E.V. 1939. Some aspects of the bank restriction period, 1797–1821. *Economic History* 4: 205–221.

Morgan, E.V. 1943. *The theory and practice of central banking, 1797–1913*. Cambridge: Cambridge University Press.

Myhrman, J. 1976. Experiences of flexible exchange rates in earlier periods: Theories, evidence, and a new view. *Scandinavian Journal of Economics* 76: 169–196.

Nachane, D.M., and N.R. Hatekar. 1995. The bullionist controversy: An empirical reappraisal. *Manchester School of Economic and Social Studies* 63: 412–425.

Officer, L.H. 2000. The bullionist controversy: A time-series analysis. *International Journal of Finance & Economics* 5: 197–209.

Ó Gráda, C. 1989. The paper pounds of 1797–1821: A co-integration analysis. Working paper. Dublin: Centre for Economic Research, University College.

Ó Gráda, C. 1991. Reassessing the irish pound report of 1804. *Bulletin of Economic Research* 43: 5–19.

Ó Gráda, C. 1993. The irish paper pound of 1797–1820: Some cliometrics of the bullionist debate. *Oxford Economic Papers* 45: 148–156.

Parnell, H. 1804. *Observations upon the state of currency in Ireland*. Dublin: M. N. Mahon.

Pressnell, L.S. 1956. *Country banking in the industrial revolution*. Oxford: Clarendon Press.

Ricardo, D. 1811. The high price of bullion. In *The works and correspondence of David Ricardo*, ed. P. Sraffa, Vol. 3, 1951. Cambridge: Cambridge University Press.

Silberling, N.J. 1924. Financial and monetary policy of Great Britain during the Napoleonic Wars. *Quarterly Journal of Economics* 38: 214–233.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper.

Bullionist Controversy

David Laidler

'Bullionist Controversy' is the label conventionally attached to the series of debates about monetary theory and policy which took place in Britain over the years 1797–1821, when the specie convertibility of Bank of England notes was suspended. The protagonists in this controversy are usually classified into two camps – 'bullionist' supporters of specie convertibility who were critics of the Bank of England, and 'anti-bullionist' adherents of an opposing viewpoint. Such labels are useful as organizing devices, but it is dangerous to apply them rigidly. The bullionist controversy was a series of debates about a variety of issues, and those debates involved a shifting cast of participants, whose views sometimes changed as controversy continued.

Although contemporary policy problems provided most of the immediate impetus for debate, the bullionist controversy was not a series of arguments about the application of well- known economic principles to a particular set of

circumstances. On the contrary, much of the debate was about fundamental questions of economic theory; and though the literature of the controversy consists largely of pamphlets, reviews, letters to newspapers, parliamentary speeches and reports, it contains contributions of crucial and lasting importance to monetary theory.

I. The Bank of England, a privately owned joint stock company, was founded in 1694 with the aim of creating a market for, and an institution to manage, the government debt arising from William Ill's participation in the wars against the France of Louis XIV. By the end of the 18th century its monopoly of note issue in the London area, and its status as the only note-issuing joint stock bank in England, had given it a pivotal position in the British monetary system. It had in fact evolved into the central bank at least of England, though not of the United Kingdom; for Ireland at this time had its own largely independent monetary system, with commercial banks operating on a reserve base provided by the Bank of Ireland in Dublin, which held its reserves in specie rather than in claims upon London. Scottish Banks too belonged to a distinct system, albeit one which held its reserves in London. Though reforms of the coinage beginning in 1696 and culminating in that supervised by Sir Isaac Newton in 1717 had been intended to create a bimetallic system, their undervaluation of silver had instead placed Britain on a *de facto* gold standard that was firmly entrenched by the last decade of the century.

By the 1790s the 'circulating medium', to use a contemporary phrase, consisted of gold coin, Bank of England and Country (i.e., non-London) Bank notes, while bills of exchange and bank deposits were widely used means of payment in wholesale transactions. Country Banks mainly held reserves on deposit with private London banks, which did not emit notes, and which in turn held reserves in the form of Bank of England liabilities. Britain's specie reserves were mainly held by the Bank of England in the form of bullion. The degree of concentration here was not as absolute as it would become later in the 19th century, but, to put it in modern parlance, Bank of England liabilities were high-powered money, and any difficulties in the banking system at large quickly put pressure on the Bank's specie reserves.

The outbreak of hostilities between Britain and Revolutionary France in 1793 precipitated just such pressure. A drain of reserves from the banking system into domestic private sector portfolios, to which the Bank of England responded by contracting its note issue, created a liquidity crisis. The crisis was alleviated by a government issue of exchequer bills, and this very fact speaks eloquently of the lack of appreciation, on the part of the Bank and Government alike, of the role and responsibilities of a Central Bank in the monetary and financial system which characterized the state of knowledge at the beginning of the bullionist controversy. Not the least of that controversy's enduring contributions was to advance understanding of these matters.

As France recovered from the political chaos associated with the Terror, and the monetary chaos created by the Assignats, the war began to go badly for Britain and her allies. By the beginning of 1797 France was clearly in the ascendant. Indeed, the completion of Bonaparte's Italian campaign at the end of that year would see only Britain remaining in the field against her. During 1795–6 the Bank of England had again attempted to counter a continuing drain of specie from its reserves by a contraction of its liabilities, and had probably thereby accentuated its difficulties. This certainly was the opinion of commentators such as Walter Boyd (1800), while Henry Thornton's (1802) analysis of the general importance of a Central Bank's standing ready to lend freely in the face of a domestic run on its reserves in order to restore and maintain confidence may be read, in part, as a criticism of the Bank of England's behaviour during this episode.

Be that as it may, by February of 1797, pressure on the Bank was again strong, and

rumours of an impending French invasion – a small force of French troops did land in Wales but was quickly captured – provoked a run on the banking system. This run began in Newcastle and quickly spread. To the Government and the Bank of England it seemed to put that institution in jeopardy, and an Order in Council of 26 February, confirmed in May by an Act of Parliament, suspended the specie convertibility of Bank of England notes. This 'temporary' suspension, initially supposed to end in June 1797, was to last until 1821. The management of an inconvertible currency – or rather partially convertible, for gold and some subsidiary silver coin continued to circulate, and during the suspension period of Bank did from time to time declare some of its small denomination notes convertible – would have been difficult enough in peacetime; but down to 1815 the Bank of England's task was frequently complicated by the need to make large transfers abroad to subsidise allies and support British forces fighting on the Continent, not to mention the disruptive effects of the Napoleonic 'Continental System' on British trade.

The body of economic analysis which a modern economist would deploy in dealing with these matters was not available in Regency Britain. The Cantillon (1734)–Hume (1752) version of the quantity theory of money, and its associated analysis of the price-specie flow mechanism was well enough known; but that dealt with a commodity money system, not with one dominated by banks, in which a large proportion of the 'circulating medium' consisted of bank notes and deposits (or cheques drawn upon them), not to mention various commercial bills. The *Wealth of Nations* (Smith 1776) contained extensive discussions of banking, but those discussions, as Checkland (1975) has argued, were largely based on Scottish oral tradition; they therefore dealt with the competitive operations of commercial banks against the background of specie convertibility and had next to nothing to say about central banking.

Much available knowledge about the operation of inconvertible paper systems was of a practical nature. It drew on the French experience with John Law's scheme, and later the Assignats, on many North American experiments before, during and after the American War of Independence, and, to a lesser extent, the 18th-century experience of Russia and Sweden with paper money. Though the Swedish experience had generated controversy which in many respects anticipated the British bullionist debate, as Eagly (1968) has shown, there seems to be no evidence that the Swedish literature was known in Britain, even to those who, like Henry Thornton, were aware of the events that had generated it.

In short, by the 1790s, institutional developments in the British monetary system had run far ahead of systematic knowledge of what we would now call the theory of money and banking. The difficulties of the suspension period focused attention on this fact, and the analysis developed during the course of the bullionist controversy had to solve fundamental problems in monetary theory as well as cope with contemporary policy issues. It is because it dealt with the first of these tasks with such success that the controversy is of enduring importance to monetary economists, and not just to historians of economic thought and economic historians.

II. The 18th-century experiences with inconvertible paper referred to above were, with few exceptions, unhappy, and it is scarcely surprising that, at the very outset, opponents of restriction in Britain warned of dire inflationary consequences. However, it was not until 1800 that rising prices, a decline in the value of Bank of England paper in terms of bullion, and an associated depreciation of the sterling exchange rate on Hamburg gave warning that all was not well. (We need not concern ourselves here with the complications caused by the fact that Hamburg was on a silver and not a gold standard.) These events generated a flurry of pamphlets, and it is generally agreed that Walter Boyd's (1800) *Letter to ... William Pitt* was the most noteworthy of these. It

stated a simple version of what was to become known as the bullionist position, namely that the suspension of convertibility had permitted the Bank of England unduly to expand its note issue and that overexpansion had in turn brought about the above-mentioned interrelated consequence.

The fact that agricultural prices had risen considerably more than the value of bullion made it possible for defenders of the Bank of England, such as Sir Francis Baring, to argue that the problem lay elsewhere than in the banking system *per se*. The Bank's defenders also raised at this early stage of the debate what was to become an important bone of contention in later monetary debates, namely the possibility that the Country Banks, by varying their note issue, could and indeed did exert an influence on the behaviour of the price level independently of the Bank of England. The preliminary 'skirmish' of 1800–1802 as Fetter (1965) called it was indecisive, but it produced Henry Thornton's *Paper Credit* ... (1802), an extraordinary treatise which systematically expounds the intellectual basis of what Viner (1937) termed the 'moderate bullionist' position in subsequent discussions.

Von Hayek suggests in his introduction to *Paper Credit* that Thornton may have been working on it as early as 1796, but in its published form, this book was a defence, albeit a constructively critical defence, of the Bank of England's policy during the early years of restriction. It was published during a lull in the debate, and its *direct* influence on the course of the bullionist controversy was therefore minor. During the 19th century the work dropped from sight, and its true stature was not thereafter widely appreciated until the appearance of von Hayek's (1939) edition. Indirectly, however, *Paper Credit* was of the first order of importance. Its author was an influential member both of the Committee of the House of Commons that investigated Irish currency issues in 1804 – see Fetter (1955) on this episode – and of the so-called Bullion Committee itself, whose 1810 report marked the high point of the controversy. Moreover, the chairman of the latter committee, Francis Horner, who, with help from Thornton and William Huskisson, was the principal author of its Report, had devoted a long and favourable review article to *Paper Credit*.in the first issue of the *Edinburgh Review*.

III. The immediate cause of the renewed controversy that led to the setting up by Parliament of the *Select Committee on The High Price of Gold Bullion* in February 1810 was a reemergence of inflationary pressures in early 1809, whose most noticeable symptoms to observers not equipped with even the concept of a price index, let alone a serviceable example of such a device, were a declining exchange rate for sterling and marked rise in the price of specie in terms of Bank of England notes. Both of these symptoms were more marked than they had been in 1800–1802, but the positions taken up in the controversy that preceded the committee's formation and accompanied its deliberations were very much those established in the preliminary skirmish of those years.

What Viner (1937) terms the 'extreme bullionist' position had been stated by John Wheatley as early as 1803, and was subsequently maintained by him. David Ricardo, whose contributions to the *Morning Chronicle* in 1809 represent his first published work in economics also argued this position, though a little more flexibly than Wheatley, notably in his (1810–11) essay on *The High Price of Gold Bullion*. Simply put, the extreme bullionist position was that the decline in the exchanges, and the increase in the price of bullion, were solely due to an excessive issue of Bank of England notes, an excessive issue which could not have taken place under convertibility. Against such views, the anti-bullionist defenders of the Bank argued that the decline in the exchanges was due to pressures exerted by extraordinary wartime foreign remittances and had nothing to do with the Bank's domestic policy. Moreover, they argued, because the Bank confined itself to making loans on the security of high quality commercial bills,

drawn to finance goods in the course of production and distribution, it was impossible that its note issue could be excessive and could cause prices to rise. The first of these arguments deals with what we would now call the 'transfer problem' and the second is a statement of the infamous *Real Bills Doctrine*.

At the outset of the bullionist controversy there existed little in the way of coherent analysis of the transfer problem under conditions of convertibility, let alone of inconvertibility. Adam Smith (1776) had stated that foreign remittances would in fact be effected by a transfer of goods rather than specie abroad, but had not explained how, while during the bullionist controversy the directors of the Bank of England consistently argued that any transfer must initially involve an outflow of specie equal in amount to the transfer itself. This position was not far removed from the naïve mercantilist analysis which Hume had so effectively attacked in 1752, and was, as Fetter (1965) has noted, quite inconsistent with the actual behaviour of the Bank's specie reserves during the French wars.

A key contributor to the analysis of the transfer problem was Thornton, and the influence of ideas first expounded in *Paper Credit* is quite evident in the Committee's *Bullion Report* of 1810 (Cannan 1919). He had shown in *Paper Credit* how a transfer of goods would be brought about under a convertible currency as a result of monetary contraction in the country making the transfer and expansion in the recipient country, and had stressed income effects as well as price level changes as critical links in the mechanism. Though he did not distinguish clearly between a convertible and an inconvertible currency, he also argued that, under post-1797 arrangements (which because of the continued circulation of gold coin did not amount to a clear-cut inconvertible system), the mechanisms in question would lead to a temporary exchange rate depreciation, even if domestic policy was such as to promote what we would now term domestic price level stability. The limits to the possible depreciation here would be set by the costs of evading legal prohibitions on the melting and export of coin.

In 1802 this analysis had formed part of Thornton's defence of Bank of England policy against bullionist critics, and it was further refined in the course of the deliberations of the Parliamentary Committee of 1803 which investigated the depreciation of the Irish pound, and on which Thornton served. At least two authors, John Hill and J.C. Herries (both anti-bullionists) were later to supplement it with the observation that a temporary depreciation created scope for short-term capital movements to help in making a transfer effective.

By 1810–11, the view that transfers could temporarily depress the exchanges under conditions of inconvertibility, and a growing scarcity of gold coin had moved the system much closer to such conditions than it had been a decade earlier, set the analysis of moderate bullionists, including Thomas R. Malthus, and of course the Bullion Committee itself, apart from that of Ricardo and Wheatley, who denied that even a temporary exchange rate depreciation could take place in the absence of a simultaneous excessive issue of domestic paper. Either this latter argument involves an implicit definition of 'excessive' and is circular; or, as Viner has suggested, it is erroneous and provides an unfortunate example of the 'Ricardian vice' of giving answers relevant to the long run equilibrium outcome of particular situations to questions having to do with the intermediate stages whereby long run equilibrium is achieved.

Disagreement among the bullionists was about temporary effects, however. Moderate bullionists were in complete agreement with their more extreme colleagues that an apparently permanent exchange depreciation could not be put down to the effects of once and for all transfers. Their view, as expressed in the 1810 *Report,* was that sterling's initial depreciation had probably been the consequence of foreign remittances, and of the effects of the Continental System on trade, but that its subsequent failure to recover was caused by an

overissue of paper money by the Bank of England. They thus rejected the Bank of England's claim that it was powerless to affect the purchasing power of paper money so long as it confined its issues to those called forth by the supply for discount of good quality bills of exchange.

The analysis of the Real Bills Doctrine set out in the *Bullion Report* is in all its essentials the same as that to be found in *Paper Credit*, and is marked by a careful discussion of the mechanisms whereby the policies espoused by the Bank could lead to overissue. In this respect it is superior to that of Ricardo, who in his essay of 1810–11, without going into any details about the processes whereby the economy might move from one long run equilibrium to another, concentrated on giving an exceptionally clear statement of the nature of the long run equilibrium relationship that rules between the quantity of paper money, the exchange rate and the price of specie (which, as Hollander (1979) persuasively argues, is to be understood in this context as standing as a proxy for what we would now term the general price level).

The Real Bills Doctrine is attributable to Adam Smith (1776) but in his work it appears mainly as a rule of behaviour for the individual commercial bank operating in a competitive system against a background of specie convertibility. To discount only good short-term bills is not perhaps bad practice for such an institution if it wishes to secure its long-term viability. To claim such a principle to be a sufficient guarantee of price level stability if adopted by a Central Bank managing something akin to an inconvertible paper currency is another thing altogether, but that is what the directors of the Bank of England did, giving to the Bullion Committee what Bagehot (1874) was later to term 'answers almost classical by their nonsense' when questioned on this matter. Adherence to the Real Bills Fallacy was by no means confined to the Bank of England. It had many defenders and even so able an economist as Robert Torrens espoused the doctrine during the bullionist controversy, though in later debates he was to be one of its most vigorous opponents. Moreover, despite its definitive refutation by Thornton and the Bullion Committee, this doctrine was to reassert itself with great regularity throughout the 19th century, and into the 20th, as Mints (1945) in particular has so carefully documented.

The critical flaw in the Real Bills Doctrine arises from its implicitly treating the nominal quantity of bills of exchange offered for discount as being determined, independently of the policies of the banking system, by the real volume of goods under production in the economy, rather than by the perceived profitability of engaging in production and trade. The latter, as Thornton, the Bullion Committee and all subsequent critics of the doctrine have pointed out, depends upon the relationship between the rate of interest at which the banking system stands ready to lend, and the rate of return that borrowers expect to earn. To put it in the language of Knut Wicksell (1898), whose analysis of these matters closely follows Thornton – even though he appears to have been unaware of *Paper Credit* – everything depends on the relationship between the 'money rate of interest' and the 'natural rate of interest'.

As the Bullion Committee argued, with the rate of interest at which banks would lend set below the anticipated rate of profit, the potential supply of bills for discount would be without limit. Under specie convertibility, a banking system that had fixed its lending rate too low would find the associated expansion of money causing a drain of reserves and the central bank would be forced to raise its lending rate. Without the crucial check of convertibility, prices and the money supply would begin to rise, as would the nominal value of new bills of exchange offered for discount in a self-justifying inflationary spiral. The Real Bills Doctrine, a relatively harmless precept under specie convertibility, thus becomes, under inconvertibility, a recipe for unlimited inflation and exchange depreciation. This conclusion is of enduring importance and is

perhaps the most significant result that emerged from the bullionist controversy.

The Real Bills Doctrine was particularly dangerous in the circumstances of 1810. The then current usury laws set an upper limit of 5 per cent to the rate of interest at which loans could be made, and the ability of the public to convert paper money into gold coin, and then melt the latter for export, an illegal but seemingly widely practised check on overissue in the earlier days of the suspension, had become less effective by 1810 as gold coin had become scarce. Moreover, what we would now term inflationary expectations had begun to become established in the business community. Though the point was not raised explicitly in the *Bullion Report,* in a parliamentary speech of 1811 on the *Report,* Thornton showed himself well aware of the implications of this for the relationship between nominal and real interest rates and the inflationary process, thus anticipating the insights of Irving Fisher (1896) by 85 years.

In placing the blame for the persistence of sterling's depreciation on the Bank of England, the Bullion Committee also took the position that the Country Banks' note issue had not exerted a major independent influence on prices. Their *Report* contained nothing approaching a formal analysis of what we would nowadays term the 'bank credit multiplier'; such analysis did not appear until the early 1820s, when it was first developed by Thomas Joplin and James Pennington, and indeed it was not widely understood until well into the 20th century. The Committee nevertheless took the position that the Country Banks' note issue, not to mention the other privately emitted components of the circulating medium, tended to expand and contract in rough harmony with Bank of England liabilities. This is a point of some interest, since in the debates of the 1830s and 1840s, the Currency School, who in their opposition to the Real Bills Doctrine were the intellectual heirs to the bullionists, took a diametrically opposite view of the significance of the Country Bank note issue and were eventually successful in having it suppressed.

In matters of monetary theory and the diagnosis of contemporary problems it is hard to fault the Bullion Committee even today. No other discussion of economic policy issues prepared by working politicians has had so sound an intellectual basis and has stood the test of time so well. It is more difficult to praise the *Report's* key policy proposal, however. So worried were its authors about sterling's depreciation, and about the capacity of the Bank of England to conduct policy competently, that in the midst of major war, and at a time when sterling had significantly depreciated, they recommended a return to specie convertibility at the prewar parity within two years. The *Bullion Report* was laid before the House of Commons in May 1811 where debate on its substance was organized around a series of resolutions and counter-resolutions. Though the Commons rejected the whole *Report* it is not without interest that the specific proposal to resume convertibility within two years failed by a significantly larger majority than did any other. It should be noted though, that in rejecting the Bullion Committee's recommendations, the House of Commons simultaneously supported resumption once peace was re-established.

IV. Subsequent experience was to prove the Bullion Committee's fears of future Bank of England profligacy unfounded. Whatever the Bank's directors may have said about their operating procedures, they clearly relied on more than a real bills rule, and, as commentators from Bagehot on have noted, their policy was, if judged by results, reasonably responsible, particularly after 1810, which saw the peak of wartime inflationary pressures. Thus debate about monetary issues had died down by 1812, but that year saw the crucial defeat of Napoleon's army in Russia. The decline in his fortunes thereafter, leading to his final surrender in 1815, set the stage for the next phase of

the bullionist controversy. This dealt mainly with the problems of implementing resumption, though the first decisive peacetime monetary measure, taken by Parliament in 1816, was to remove the legal ambiguity which had persisted since 1717 about the status of silver in Britain's monetary system by formally placing the country on a gold standard, albeit one in which convertibility was still suspended.

The end of a war that had lasted for more than two decades was inevitably an occasion for considerable economic dislocation. Agriculture and metalworking industries in particular suffered badly from the re-establishment of peacetime patterns of production and trade. A simultaneous general fall of prices in terms of gold, upon which was superimposed a contraction of Bank of England liabilities and therefore an approach of sterling to its prewar parity, was associated with widespread distress. In such circumstances, it is hardly surprising that there was much political opposition to early resumption. By and large, this opposition was not grounded in any coherent economic analysis, except in Birmingham. In this city, the centre of the metalworking industries, opposition to resumption was articulated by Thomas and Matthias Attwood and their associates, and the Birmingham School showed a keen appreciation of the effects of monetary contraction and deflation upon employment, and an understanding that an appropriately managed monetary system based on inconvertible paper might, in principle, be a viable method of avoiding such problems.

At their best the Birmingham School anticipated Keynesian insights of the 1930s, but their analysis often degenerated into crude inflationism, particularly in their later writings. In any event, they were always a small minority among those whom we would nowadays recognize as economists. The vast majority of these always supported the principle of resumption at the 1797 parity. The value of Bank of England paper in terms of gold was either regarded as a good measure of its purchasing power over goods in general, or stability in the good value of money was looked upon as 'natural' and desirable in its own right; and there was widespread agreement that wartime inflation had been unjust to creditors. The problems of those who had incurred debts during the war, after paper had depreciated, provided some of the impetus to popular opposition to resumption immediately after the war, particularly in agricultural areas, but it is nevertheless fair to argue that a curious moral one-sidedness about the redistributive effects of inflation emerged among the majority of economists during this stage of the bullionist controversy. This one-sidedness, which perhaps had its roots in Hume's view of credit markets in which the typical borrower is an improvident consumer and the typical lender a frugal producer, has played an important role in debates about inflation ever since.

If there was, then, wide agreement about the ultimate desirability of resuming convertibility at the 1797 parity, its advocacy was nevertheless tempered with caution after 1812. In contrast to the *Bullion Report's* unconcern about such matters, later discussion did pay attention to the potentially disruptive effects on output and employment of the deflation needed to implement it. Two problems were recognized: first, deflation was needed to restore sterling to its old parity with gold; and, second, there was the possibility that the increased demand for gold implied by a resumption of convertibility might itself create more deflation by driving up the relative price of specie. The end of the war was, as we have already noted, the occasion for significant price level falls, both in terms of gold, but even more in terms of Bank of England paper, whose quantity in circulation contracted considerably. The latter contraction was not, according to Fetter (1965), the result of any conscious policy decision on the part of the Bank of England, but it did have the effect of weakening any practical case

against resumption by reducing the amount of further deflation needed to implement it.

Ricardo dominated the later stages of the bullionist controversy, as Thornton had dominated its earlier stages, and he is often regarded as having been unconcerned about deflation. Such unconcern would be consistent with the Ricardian vice of underplaying the importance of the short run in economic life, but, as Hollander (1979) has shown, this view of his position is not sustainable. Ricardo's 1816 *Proposals for An Economical and Secure Currency* were motivated by a desire to mitigate further deflation as well as by a desire to put the British monetary system upon an intellectually sound basis. He argued that, with resumption, Britain adopt a paper currency rather than one with a high proportion of gold coin, and that the Bank of England should hold against it a reserve of gold ingots in terms of which notes could be redeemed. One practical advantage of this scheme was that by economizing on gold, it would put little upward pressure on its value when it was implemented, and Ricardo pointed out this advantage. He mainly justified his proposal in more general terms, though, stressing the desirability *per se* of economizing on scarce precious metals when paper would serve equally well as currency, an argument which harked back to Adam Smith's defence of paper money in the *Wealth of Nations*.

Ricardo's ingot plan was adopted in 1819 by Parliament, of which he was by then a member, as a basis for resumption; but second thoughts about it soon set it, for quite practical reasons. Counterfeiting of bank notes had been virtually unknown before 1797, but the increased circulation of low denomination Bank of England notes thereafter had offered considerable temptation to forgers. The years 1797–1817 saw over 300 capital convictions for the offence. These convictions and, as Fetter (1965) records, the fact that clemency seems to have been granted or refused on the recommendation of the Bank, brought much opprobrium upon that institution from a public among which opposition to the widespread use of capital punishment was becoming intense. A paper currency backed by gold ingots might have been economical and secure, but it did not remove the temptation to forgery. Hence Ricardo's ingot plan was dropped, and when resumption was finally implemented in 1821, gold coins replaced small denomination notes in circulation. Ricardo's ingot plan was not forgotten, however; it was to be the starting point of Alfred Marshall's symmetallic proposals of (1887), and something very like it was implemented in Britain in 1925 when the country once again resumed gold convertibility in the wake of a wartime suspension. The similarities here were no accident. The literature of the bullionist controversy, not least Ricardo's contributions to it, was much read and cited throughout the 19th century and into the 20th, not least by participants in the monetary debates of the 1920s.

V. The resumption of 1821 was not the unmitigated disaster that the 1925 return to gold was to be, not least because the amount of deflation needed after 1819 to make the 1797 parity effective was rather minor. Nevertheless, resumption did not put an end either to monetary problems or to debate. Even the rather small amount of deflation needed after 1819 was hard for the economy to digest, and a fitful recovery thereafter ended, in 1825, in the first of a series of financial crises that were to recur at roughly decennial intervals for the next half century. Thus, if 1821 marked the end of the bullionist controversy, it also marked the beginning of a new period of debate about the monetary system, and in particular about the conduct of monetary policy and the design of monetary institutions under a gold standard. This debate would, in due course, culminate in a second famous controversy, that between the Currency School and the Banking School.

There is considerable continuity between these later debates and the bullionist controversy, and this simple fact attests to the important contributions which were made during its course. In only a quarter century, 18th-century analysis of commodity money mechanisms had been adapted to the circumstances of a

modern banking system, and the monetary economics of the open economy under fixed and flexible exchange rates had taken on a form that is recognizable even today. Moreover, the foundations of the theory of central banking under commodity and paper standards were also developed. It is hard to think of any other episode in the history of monetary economics when so much was accomplished in so short a period.

See Also

- Banking School, Currency School, Free Banking School
- Horner, Francis (1778–1817)
- Money, Classical Theory of
- Quantity Theory of Money
- Ricardo, David (1772–1823)
- Thornton, Henry (1760–1815)

Reference

Bagehot, W. 1874. In *Lombard street, a description of the money market*, ed. Frank C. Genovese. Granston: Richard Irwin,1962.

Boyd, W. 1800. Letter to the right honourable William Pitt on the influence of the stoppage of issue in specie at the Bank of England; On the prices of provisions, and other commodities. London.

Cannan, E. (ed.). 1919. *The paper pound of 1797–1821: The bullion report*. London: P.S. King & Son. Second (1921) edition, reprinted by Augustus M. Kelley, New York, 1969.

Cantillon, R. 1734. *Essai sur la nature du commerce en général.* Trans. and ed. Henry Higgs. London: Re-issued for the Royal Economic Society by Frank Cass & Co., 1959.

Checkland, S. 1975. Adam Smith and the bankers. In *Essays on Adam Smith*, ed. A.S. Skinner and T. Wilson. Oxford: The Clarendon Press.

Eagly, R.V. 1968. The Swedish and English bullionist controversies. In *Events ideology and economic theory*, ed. R.V. Eagly. Detroit: Wayne State University Press.

Fetter, F.W. 1955. *The Irish pound 1797–1826*. London: Allen & Unwin.

Fetter, F.W. 1965. *Development of British monetary orthodoxy 1797–1875*. Cambridge, MA: Harvard University Press.

Fisher, I. 1896. Appreciation and interest. *AEA Publications* 3(11): 331–442.

Hollander, S. 1979. *The economics of David Ricardo*. Toronto: University of Toronto Press.

Hume, D. 1752. Of money, of the balance of trade and of interest. In *Political discourses*. Edinburgh: Fleming. Subsequently incorporated in the 1758 edition of *Essays, Moral Political and Literary* London. Reprinted London: Oxford University Press, 1962.

Marshall, A. 1887. Remedies for fluctuations of general prices. *Contemporary Review* March; reprinted as ch. 8 of *Memorials of Alfred Marshall,* ed. A.C. Pigou, London: Macmillan, 1925.

Mints, L. 1945. *A history of banking theory.* Chicago: University of Chicago Press.

Ricardo, D. 1809. Contributions to the *Morning chronicle.* Reprinted in, *Works and correspondence of David Ricardo,* ed. P. Sraffa, vol. III. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1810–11. The high price of gold bullion, a proof of the depreciation of bank notes. Reprinted in *Works* . . ., ed. P. Sraffa, vol. III. Cambridge: Cambridge University Press, 1951.

Ricardo, D. 1816. Proposals for an economical and secure currency, Reprinted in *Works . . ., ed. P.* Sraffa, vol. IV. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations.* London. Reprinted in two vols, ed. R.H. Campbell, A.S. Skinner and W.B. Todd, Oxford: Clarendon Press, 1976.

Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain.* London. Edited with an Introduction by F.A. von Hayek, London: George Allen & Unwin, 1939; reprinted, New York: Augustus Kelley, 1962.

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper Bros.

Wheatley, J. 1803. *Remarks on currency and commerce*. London: Cadell and Davies.

Wicksell, K. 1898. *Interest and prices.* Trans. R.F. Kahn. London: Macmillan for the Royal Economic Society, 1936.

Bullock, Charles Jesse (1869–1941)

A. W. Coats

Bullock was born in Boston on 21 May 1869. Trained partly by correspondence course directed by R.T. Ely, he graduated from Boston University in 1892 while employed as a high school principal, a combination not uncommon at the time. Following his Wisconsin Ph.D. in

1895 he taught economics at Cornell, Williams (1899–1903) and Harvard (1903–34), where he directed the Committee on Economic Research from 1917 to 1929. While public finance was his principal field, he also made contributions to international economics, which was unusual before 1914, and the history of economics. The author of several successful textbooks, his major theoretical contribution was 'The Law of Variable Proportions' (1902). He served as adviser on taxation in Massachusetts and other states, and was president of the National Tax Association from 1917 to 1919.

Selected Works

1895. *The finances of the United States, 1775–1789, with special reference to the budget*. Madison: University of Wisconsin Press.
1897. *Introduction to the study of economics*. Boston: Silver Burdett & Co. 4th ed., revised and enlarged, 1913.
1900. *Essays on the monetary history of the United States*. New York: Macmillan Co.
1902. The variation of productive forces. *Quarterly Journal of Economics* 16:473–513.
1905. *Elements of economics*. Boston: Silver, Burdett & Co. Rev. enl. edn, 1923.
1936. *Economic essays*. Cambridge, MA: Harvard University Press.

Bunch Maps

Wilfred Corlett

Bunch maps were developed by Ragnar Frisch (1934) to deal with the problems of confluence analysis. By 'confluence analysis' he meant the study of several variables in some sets of which a regression equation might have a meaning, while in others it might not because of the existence of more than one relation between the variables. Frisch's exposition of bunch maps was based on a situation where each variable in a set could be split into two components: one, the systematic component, was connected with the other variables; the other, the disturbance, was not so connected. The method was used to try to determine sets of variables in which one, and only one, exact linear relation held between the systematic components of the variables. Examples of the use of the method were given for constructed data where exact relations did exist. It is less clear whether they were assumed to exist in examples of applications to actual economic data. The other major applications of bunch maps were in Richard Stone's work on consumers' expenditure (Stone 1945, 1954), but he did not consider an assumption of exact linear relations between systematic components as satisfactory.

In a full analysis of a number of variables, the bunch map was based on regressions calculated for every possible subset of two or more variables with minimization in the direction of every member of the subset. Each variable was normalized to give a unit sum of squares over observations of deviations from means. For any pair of variables, x_i and x_j (ij), in the subset, if the regression with minimization in the direction of x_k were written to express x_i in terms of the other variables, the coefficient of x_j would be minus the ratio of the cofactors of r_{jk} and r_{ik} in the correlation matrix of the variables in the subset. These cofactors were used as ordinate and abscissa respectively, but with one sign changed in such a way that the abscissa was positive, to obtain a point in a diagram. Similar points were plotted in the same diagram for all possible x_k in the subset and labelled with k. The points were joined to the origin to give the individual bunch with its beams. There was a separate bunch for every pair of variables in every subset. Together they formed the bunch map.

The bunch map was used mainly for comparing bunches of two subsets where the second contained the variables in the first plus an additional one. Attempts were made to classify the added variable as useful, superfluous or detrimental. Criteria which suggested that a variable was useful included the tightening of the bunch, a change in its general slope and the beam associated with the new variable being inside the bunch. The length of the beams and changes in length

were also considered. An explosion of the bunch showed that the new variable was detrimental; that is, it introduced multiple relations.

The complexity of the procedure and the apparent subjectivity of combining different criteria in classifying variables may have contributed to the relatively small impact of bunch maps on applied econometrics, despite frequent references in textbooks and other works on econometric techniques. Frisch's analysis did, however, draw more attention to the dangers of errors of measurement in multicollinear situations than is common in more recent discussions of multicollinearity.

See Also

- ▶ Econometrics
- ▶ Frisch, Ragnar Anton Kittel (1895–1973)
- ▶ Multicollinearity
- ▶ Simultaneous Equations Models

Bibliography

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

Stone, J.R.N. 1945. The analysis of market demand. *Journal of the Royal Statistical Society* 108(parts 3 and 4): 286–382.

Stone, J.R.N. 1954. *The measurement of consumers' expenditure and behaviour in the United Kingdom, 1920–1938*, vol. 1. Cambridge, UK: Cambridge University Press.

Bundling and Tying

Barry Nalebuff

Abstract

Bundling can be thought of as akin to a volume discount, but one where the volume is based on aggregate sales across products. Instead of offering a discount for buying two apples rather than one, the customer is given a better price for buying an apple and an orange together. Bundling may be used to reduce cost and improve quality, and for price discrimination. While the Chicago School has argued that a monopolist cannot gain by bundling its monopoly good with a competitive product, recent work suggests that in a dynamic game bundling can help protect and leverage market power.

Keywords

Antitrust policy; Bundle discounting; Bundling; Chicago School; Complementarities; Consumer surplus; Cournot, A.; Envelope theorem; Market power; Markup; Metering; Mixed bundling; One-monopoly profit argument; Price discrimination; Pure bundling; Two-part tariffs; Tying

JEL Classifications

L10

Bundling is a prevalent feature of pricing. It is akin to a volume discount, but where the volume is based on aggregate sales across products. Instead of offering a discount for buying two apples rather than one, the customer is given a better price for buying an apple and an orange together.

Under pure bundling, two goods A and B are sold together only as a package. Under mixed bundling, customers can also buy each good. Typically, the bundle is offered at a discount to the individual prices.

Mixed bundling is the most general case. A pure bundle can be thought of as a case where the individual prices exceed the bundle price, so that no one has an incentive to purchase anything but the bundle. Tying can also be viewed as a special case of mixed bundling; customers are offered prices for A and B together or for B alone, but not A without B.

The first to study bundling was Cournot (1838), who showed how it solves a double markup problem for complementary products. Bundling may increase efficiency more directly by improving quality and reducing cost (see Evans and Salinger 2004, 2005). Manufacturers

gain scale economies by standardizing the combination of goods and guaranteeing that the components work together.

The early bundling literature debated whether it was used to leverage market power or price discriminate. This debate was stimulated by a series of cases in which the courts viewed bundling (and tying) as anti-competitive. Director and Levi's (1956) and Stigler's (1963) influential Chicago School argument claimed that a monopolist cannot gain by leveraging its power from one market to another.

Starting with Whinston (1990), the current literature suggests that, in a dynamic setting, bundling can profitably leverage market power by deterring entry, excluding one-good rivals, and amplifying existing market power. Three review articles provide a guide to the literature and antitrust cases: Kaplow (1985), Nalebuff (2003a), and Kobayashi (2005).

The Chicago School Argument

In response to *United States v. Loew's* (1962), Stigler (1963) argued that block booking (selling movies bundled rather than individually) was best viewed as price discrimination. He argued that a monopolist in product A could not make more money by requiring buyers to take a product B that was competitively available – the alternative strategy of selling A alone for a price of p-c, where p is the bundle price and c the marginal cost of B, is more profitable. Any sale of A at price p-c would be just as profitable as selling the bundle at p. Yet anyone willing to buy the bundle at p would also be willing to buy A alone at p-c, as, by assumption, B is available at the competitive price of c. Bundling is weakly worse as it might cause the firm to lose sales to customers who value A at p-c but do not value B at c and thus do not buy the bundle. This has become known as the 'one-monopoly profit' or 'Chicago School' argument (see Director and Levi 1956; Bork 1978).

If leverage does not explain bundling, something else must. Stigler suggested price discrimination.

Price Discrimination

The idea of bundling (and tying) as price discrimination dates from Bowman (1957) and Burstein (1960). As Burstein noted, a monopolist would generally like to employ a two-part tariff in pricing. Requiring customers to buy an overpriced B is an indirect way to charge a lump-sum fee.

If the monopolist starts from a profit-maximizing price, then (by the envelope theorem) profits lost from cutting price will be very small. In contrast, existing consumers will gain a great deal, and so will be willing to buy B at a inflated price in return for a lower-priced A.

The problem is that other producers of B end up excluded. Customers of A won't switch to lower-priced B goods because they don't want to lose the discount on A. While the monopolist could have used a two-part tariff directly, such pricing schedules seem rare in practice. Bundle pricing as a two-part tariff is explored in Mathewson and Winter (1997) and Nalebuff (2004b).

Two-part pricing becomes even more effective when B's demand is correlated with A's value. This leads to metering. For example, a firm selling printers would like to charge high-value customers more. But customer valuation may be unobservable. However, if value is correlated with usage, then a per-page charge would allow the seller to charge high-value customers more. A per-page charge could be levied directly, although that would require monitoring usage. In practice, sellers patent the shape of their toner cartridge, thus requiring users to buy toner at a premium price.

These results rely on B's demand being either elastic or heterogeneous. Bundling permits price discrimination even when A and B are consumed in fixed amounts. Consider movies. If regional variation in the valuation of two movies is negatively correlated, a distributor can profit more by pricing the movies as à package than by selling them a la carte. Bundling reduces demand heterogeneity and thus captures more of consumer surplus (see Adams and Yellen 1976).

The advantages of the bundle discount strategy are remarkably general.

McAfee et al. (1989) show that, for any two goods independent in value, a firm with market

power will find an advantage offering them at a bundle discount (holding individual prices constant) – an impressive result, given the near endless opportunities for bundling products with independent values.

One intuition for their argument is that discounting via bundling leads to twice the demand expansion for the same price reduction. Consider the offer to lower A's price by one dollar if you buy B. The cost of the offer is one dollar to all customers who would have bought both A and B at the previous prices. The gain is the new demand from customers who were buying B but not A, as they now have an opportunity to get A at a dollar off. If A was priced optimally to begin with, then the incremental profit from increased demand should just offset the lost revenue on existing customers. (Here demand independence is critical, as it implies that customers buying B are representative of the entire A market.) So far, everything is a wash. However, the dollar off A if you buy B is the same as a dollar off B if you buy A. Thus there is a second set of incremental customers: those already buying A but on the margin on B. Demand for B expands without imposing any further cost in terms of lost revenue. The ability of the bundle to expand demand on two fronts for one discount is the 'special sauce' behind bundling.

Bundling to Leverage Monopoly

The recent re-examination of bundling as leveraging market power and foreclosing rivals uses dynamic reasoning, which is absent in the Chicago argument.

For example, a monopolist in A might bundle A with B to drive rivals out of the B market. The motivation could be to monopolize what was previously a competitive B market, or to protect the A monopoly. Eliminating firms in the B market protects A if being in the B market facilitates entry into A. The US Department of Justice (1998) argued thus in explaining Microsoft's motivation to bundle Explorer with Windows – defeating Netscape would prevent it from threatening Microsoft's operating system monopoly.

The first dynamic model appears in Whinston (1990), where the bundler has market power in both A and B and uses the bundle to deter potential entrants. The monopolist is concerned that there may be a rival who can produce B at a lower cost. In defence, it commits itself to sell A only along with B. Thus, if a rival were to create a lower-cost B, the monopolist would not concede, as that would cost it its profits in A sales. Since the monopolist is committed to selling A only along with B, it would have to subsidize B in order to sell A. Even more efficient B good rivals won't enter, realizing that they won't win; this preserves monopoly profits in B.

Nalebuff (2004a) offers a second perspective. Absent entry, the dual monopolist gains via price discrimination. With entry (and heterogeneous consumer preferences), the firm would rather respond with a bundle than with head-to-head competition (and thereby lose all profits in the B market).

The incumbent's bundling reduces the potential market available to the entrant. The entrant is mostly limited to those customers who like B but not A. This market may not be large enough to cover costs of entry or to achieve a minimum efficient scale (Carlton and Waldman 2002).

The bundling models illustrate the challenge for anyone contemplating entry against Microsoft Office. Given the large discount for buying the Office bundle, a firm that developed a better word-processing program (and nothing else) would find its market limited to those who value word processing, but not spreadsheets or presentations. The entrant could try to sell to those who already have Word, but that would limit the price to its product's incremental value over Word, which is much less than what it can charge customers who don't already have Word.

A firm could always develop a rival bundle of products. But this also discourages entry, as it is much harder to develop two better products than one. Furthermore, it turns out that bundle-against-bundle competition is particularly fierce (see Matutes and Regibeau 1992).

These examples of bundling emphasized the use of pure bundles as a way of protecting and leveraging market power. Even with mixed bundling, firms can achieve similar results by keeping the component prices artificially high.

A bundle discount may be large due to a low bundle price or high individual prices, prices that might exceed monopoly levels. Although entry is blocked in both cases, the welfare implications are different, as discussed in Greenlee, Reitman and Sibley (2004). Bundling can be used to create a horizontal price squeeze, an issue considered by the Supreme Court in *Ortho Diagnostic v. Abbott Lab.* (1996) and developed in Nalebuff (2005).

A bundle discount leads to foreclosure if even the *monopolist* could not afford to sell B at a large enough discount to offset the loss of the bundle discount. Exclusionary bundling arises when the incremental price for an A–B bundle over A alone is less than the long-run average variable costs of B.

Bundling Complements

An incentive to bundle arises when two products are perfect complements, so that customers care only about their combined price. Cournot (1838) considered copper and zinc, which combine to produce brass; a more modern example would be hardware and software.

Two monopolists selling A and B independently will charge inefficiently high prices. Were the two firms to merge or coordinate their pricing, they can lower prices and raise profits. The gain from bundling complements is the horizontal equivalent of vertical integration to avoid double marginalization. As consumers and firms are both better off, this is a Pareto improvement.

The situation is more complicated if there are multiple producers of A and B. Nalebuff (2000) and Choi (2001) consider the case where two firms are able to solve the coordination problem while their rivals are not. This issue arose in 2001 when the European Commission blocked the proposed US$42 billion merger between General Electric and Honeywell. The Commission was concerned that the merger would allow the combined firm to better coordinate the pricing of airplane engines and avionics, and give it an advantage over engine-only rivals such as Rolls Royce or avionics-only rivals such as Thales or Rockwell Collins (see Nalebuff 2003b, for a cautionary note).

Bundling can change competition in two ways. When a bundle competes against components, the bundled seller is better able to coordinate pricing and gains share against his rivals. Profits may not rise as rivals respond to their reduced market share with lower prices. When there is bundle-against-bundle competition, as shown by Matutes and Regibeau (1992), prices are the lowest of all, and profits fall substantially. Customers benefit from the lower prices but lose the ability to mix and match and thereby buy their ideal mix of product.

There may also be a combination of these two effects. With an imbalance between A and B producers, only some firms are able to offer bundles, and these firms compete aggressively. The left-out firms have only one good and end up disadvantaged; see Gans and King's (2004) analysis of bundle discounts offered by supermarket and gasoline retailers in Australia.

Conclusions

There is no grand unification theory of bundling. The decision to bundle is connected both to product design and to pricing. While price discounts are typically pro-competitive, in some cases bundling creates a cause for antitrust concern as it can be used to protect and leverage market power.

See Also

► Price Discrimination (Theory)

Bibliography

Adams, W., and J. Yellen. 1976. Commodity bundling and the burden of monopoly. *Quarterly Journal of Economics* 90: 475–498.

Bork, R. 1978. *The antitrust paradox: A policy at war with itself*. New York: Basic Books.

Bowman, W. 1957. Tying arrangements and the leverage problem. *Yale Law Journal* 67: 19–36.

Burstein, M. 1960. The economics of tie-in sales. *Review of Economics and Statistics* 42: 68–73.

Carlton, D., and M. Waldman. 2002. The strategic use of tying to preserve and create market power in evolving industries. *RAND Journal of Economics* 33: 194–220.

Choi, J. 2001. A theory of mixed bundling applied to the GE-Honeywell merger. *Antitrust Magazine* 16: 32–33.
Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses. Paris: Hachette.* Trans. N. Bacon as Researches into the Mathematical Principles of the Theory of Wealth. Mountain Center, CA: James and Gordon. 1995.
Director, A., and E. Levi. 1956. Law and the future: Trade regulation. *Northwestern University Law Review* 51: 281–296.
Evans, D., and Salinger, M. 2004. *An empirical analysis of bundling and tying: Over-the-counter pain relief and cold medicines*. Working Paper No. 1297. Munich: CESifo.
Evans, D., and M. Salinger. 2005. Why do firms bundle and tie? Evidence from competitive markets and implications for tying law. *Yale Journal on Regulation* 22: 37–89.
Gans, J., and S. King. 2004. Supermarkets and shopper dockets: The Australian experience. *Australian Economic Review* 37: 311–316.
Greenlee, P., Reitman, D., and Sibley, D. 2004. *An antitrust analysis of bundled loyalty discounts*. Discussion Paper No. 04–13. Economic Analysis Group, Antitrust Division, US Department of Justice.
Kaplow, L. 1985. Extension of monopoly power through leverage. *Columbia Law Review* 85: 515–554.
Kobayashi, B. 2005. Does economics provide a reliable guide to regulating commodity bundling by firms? A survey of the economic literature. *Journal of Competition Law and Economics* 1: 707–746.
Matutes, C., and P. Regibeau. 1992. Compatibility and bundling of complementary goods in a duopoly. *Journal of Industrial Economics* 40: 37–54.
Mathewson, F., and R. Winter. 1997. Tying as a response to demand uncertainty. *RAND Journal of Economics* 28: 566–583.
McAfee, R., J. McMillan, and M. Whinston. 1989. Multiproduct monopoly, commodity bundling, and correlation of values. *Quarterly Journal of Economics* 104: 371–384.
Nalebuff, B. 2000. Competing against bundles. In *Incentives, organization, and public economics*, ed. P. Hammond and G. Myles. Oxford: Oxford University Press.
Nalebuff, B. 2003a. *Bundling, tying, and portfolio effects*. DTI Economics Papers No. 1. Online. Available at http://www.dti.gov.uk/files/file14774.pdf and http://www.dti.gov.uk/files/file14775.pdf. Accessed 4 July 2006.
Nalebuff, B. 2003b. Bundling and the GE–Honeywell merger. In *The antitrust revolution*, 4th ed., ed. J. Kwoka and L. White. New York: Oxford University Press.
Nalebuff, B. 2004a. Bundling as an entry barrier. *Quarterly Journal of Economics* 119: 159–187.
Nalebuff, B. 2004b. *Bundling as a way to leverage monopoly.* Working Paper No. ES-36, Yale School of Management.
Nalebuff, B. 2005. Exclusionary bundling. *Antitrust Bulletin* 50: 321–370.
Ortho Diagnostic Sys. v. Abbott Lab., 920 F. Supp. 455 (*S.D.N.Y.* 1996).
Stigler, G. 1963. US v. Loew's Inc.: A note on block booking. *1963 Supreme Court* Review (1964), 152–7. Reprinted in G. Stigler, ed., *The Organization of Industry.* Chicago: University of Chicago Press, 1968.
United States v. Loew's Inc. 371 U.S. 38 (1962).
US Department of Justice. 1998. *The United States v. Microsoft Corporation.* Civil Action No. 98–1232. Online. Available at http://www.usdoj.gov/atr/cases/f1700/1763.htm. Accessed 4 July 2006.
Whinston, M. 1990. Tying foreclosure, and exclusion. *American Economic Review* 80: 837–859.

B

Burden of the Debt

Robert Eisner

Public debt constitutes private assets. The deficit of one sector of the economy is the surplus of another.

Thus, for a closed economy, internally held public debt is not the obvious burden it is to an individual. If the private sector pays taxes to service the debt, it is also the private sector which receives the proceeds of these taxes in payments of interest and principal. If taxes could be lumpsum, with no marginal effect upon economic behaviour, and were anticipated with certainty, and if public and private borrowing costs were the same, it could be argued that the public debt would be irrelevant, except for distributional effects.

This would imply that it would make no difference whether public expenditures were financed by current taxes or by borrowing (which would create a public debt that would be serviced by future taxes). This proposition, considered but dismissed by Ricardo, was labelled by Buchanan (1976; see also 1958) the Ricardo Equivalence Theorem after being refurbished by Barro (1974).

The public debt is not considered neutral, even with lump-sum taxation, if only because people are mortal. Those currently holding debt and receiving interest will escape some taxation by death. Barro's answer was to postulate agents with preference functions in which the assets and liabilities of

their descendants were arguments. Hence the current generation's holdings of public debt in excess of the present value of their own consequent tax liabilities would be matched by their need to adjust their bequests to leave their heirs uninjured by future taxes necessary to service the debt.

To this there are many objections, including such obvious ones as the fact that some current agents have no heirs, that others do not care about their heirs, and that still others are at 'corner solutions', so that the amount that they give to (or receive from) their children will not be affected. There are further objections in terms of uncertainty as to life span, both for current agents and their children, and even with regard to the number of their heirs and of their heirs' heirs for whom provision should be made. These objections to the effective assumption of immortality, along with differences in public and private borrowing costs, and of course the fact that most taxation is not and cannot reasonably be expected to be of a lump-sum variety, has led to considerable rejection of the equivalence theorem (see Buiter and Tobin 1979).

Whether (and if so, in what way) the public debt is a burden then becomes a highly conditional issue. While most theoretical discussions have apparently accepted the premise of full, market-clearing equilibrium, the more relevant circumstance is frequently one of underemployment related to insufficient aggregate demand. In this situation, public debt, far from being a burden, is likely to induce greater consumption, as is made particularly clear by Modigliani's life-cycle hypothesis (Modigliani and Brumberg 1954; Ando and Modigliani 1963). Those with greater wealth, in the form of public debt or other assets, will consume more now and plan to consume more in the future as well. Within a framework of rational expectations (without market-clearing), firms should then complement the increased consumption with increased investment to meet current and future consumption demand. Current output and employment would thus be higher, and a greater capital stock would be available for the future.

The existence of public debt, including non-interest-bearing debt in the form of government money, also facilitates inter-generational contracts. It permits the current working generation to save and exercise a claim for retirement support from the next generation in the absence of the ability to accumulate non-depreciating capital.

This possible benefit of widening available choices to saving and consumption leads some to view public debt as a burden. For if the public debt increases current consumption, it is argued that there must be less saving and hence a lesser accumulation of capital. Public debt proves a replacement for assets in the form of productive capital. The economy then suffers a lesser capital stock and hence less production, and, in equilibrium, less consumption as well. This argument has been extended by Feldstein (1974) to implicit government debt in the form of 'social security' or pension commitments.

While this argument, as indicated above, is clearly reversed in a situation of underemployment, where added consumption is likely to mean added investment as well, its macroeconomic applicability even in a condition of full-employment equilibrium, is questionable. For an increase in the public debt, in an economy already in full-employment equilibrium, would generate excess demand which would raise prices. If government non-interest-bearing debt in the form of money were increased in proportion to the increase in interest-bearing debt, the economy could then move to a new equilibrium in which prices would be higher but the *real* value of the public debt, the real value of the quantity of money, the rate of interest and all other real variables, including the rates of investment and consumption would be unchanged. If, again under conditions of full employment, the government imposes a permanent nominal deficit on a no-growth economy, it will generate a rate of inflation corresponding to the rate of increase of nominal debt. Hence real debt will not rise, and once more none of the presumed burden of increased debt will develop.

This suggests the existence of considerable confusion between real and nominal magnitudes. It is essentially the real public debt that matters. The nominal, par value of public debt has risen in many countries while rising interest rates and rising prices have caused substantial declines in

its real, market values. It is hence important to correct measures of budget surpluses and deficits so that they correspond to *real* changes in the public debt. Suppose, for example, a nominal deficit of $100 billion and interest and price effects on the real value of an outstanding public debt of $2000 billion such that its real, market value, aside from the current deficit, declines to $1850 billion. In relevant, real terms the budget may then be viewed as in *surplus* by $50 billion, which is equal to the $150 billion 'capital gain' or 'inflation tax', minus the $100 billion nominal deficit (see Eisner and Pieper 1984; Eisner 1986).

Whether a burden or a blessing, public debt may, along these lines, be better evaluated in terms of its relation to the income or product of the economy. The debt may thus be viewed as rising, in a relative sense, only when it increases more rapidly than gross national product. In an economy with a debt-to-GNP ratio of 0.5, for example, this would mean that with a growth rate of, say, 8 per cent per annum (consisting, approximately, of 3 per cent real growth and 5 per cent inflation), the debt could grow at 8 per cent per year, implying a deficit equal to 4 per cent of GNP, with no change in the ratio of debt to GNP. A corollary of this is that, in a growing economy, there is always some equilibrium debt-to-GNP ratio consistent with any deficit-to-GNP ratio, that is

$$[\text{Debt}/\text{GNP} = (\text{DEF}/\text{GNP}) \div (\Delta\text{GNP}/\text{GNP})].$$

If public debt is related to public assets, financial and tangible, the net public debt is likely to prove considerably less than the gross public debt, and the net worth or net assets of the public sector are likely to prove positive even in economies with large public debts. In a larger sense, public debt may well be related to total wealth of the economy, private as well as public, and human as well as non-human. A larger public debt may then be associated with greater public wealth. The public debt may properly be viewed as a burden on the economy, however, to the extent that it diminishes total real wealth. It may do so, if it does not increase public capital, by reducing the supply of private capital and/or the supply of labour.

That public debt would reduce the supply of private capital is, as already pointed out, questionable. With regard to the supply of labour, the force of the argument would depend upon agents finding their wealth in the form of public debt so great that their supply of labour to secure additional income or wealth would be significantly curtailed. The real magnitudes of public debt, or ratios of public debt to gross national product, are nowhere sufficient to make this a serious concern. In the United States, for example, interest payments on the federal debt in 1986, despite half a decade of presumably huge deficits, represent no more than 3 per cent of gross national product. The real interest received by bondholders, after adjusting for the inflation loss in the principal of their bonds, is less than 2 per cent of gross national product. The public debt would have to be many times as large before private income from holding of the debt would be sufficient to have an appreciable effect in reducing the supply of labour (or other factors of production). Indeed, it may not be *possible* for the real debt to be sufficiently high to impinge significantly upon the supply of labour. For the necessary increases in nominal debt would generate such excess demand and consequent increases in the price level that an upper bound to the real debt would be reached before its effects upon supply could be significant.

All of this relates to internally or domestically held public debt. Public debt held by other countries or their nationals is another matter. If that debt is denominated in a country's own currency, it too can always be paid off by money creation and depreciated by inflation. If there is an external debt in foreign currencies, however, there is a real burden, which can, if the debt is sufficiently large, prove overwhelming. In the case of such external debt, this burden must be carefully balanced against any benefits in terms of income from the wealth or assets which the debt may have financed.

See Also

- ▶ Crowding Out
- ▶ National Debt
- ▶ Public Debt
- ▶ Ricardian Equivalence Theorem

Bibliography

Ando, A.K., and F. Modigliani. 1963. The 'life cycle' hypothesis of saving: Aggregate implications and tests. *American Economic Review* 53: 55–84.

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Buchanan, J.M. 1958. *Public principles of public debt*. Homewood: Irwin.

Buchanan, J.M. 1976. Barro on the Ricardian equivalence theorem. *Journal of Political Economy* 84: 337–342.

Buiter, W.H., and J. Tobin. 1979. Debt neutrality: A brief review of doctrine and evidence. In *Social security versus private saving*, ed. George M. von Furstenburg. Cambridge, MA: Ballinger.

Eisner, R. 1986. *How real is the federal deficit?* New York: Free Press, Macmillan.

Eisner, R., and P.J. Pieper. 1984. A new view of the federal debt and budget deficits. *The American Economic Review* 74: 11–29.

Feldstein, M. 1974. Social security, induced retirement, and aggregate accumulation. *Journal of Political Economy* 82: 905–925.

Modigliani, F., and R. Brumberg. 1954. Utility analysis and the consumption function: An interpretation of cross-section data. In *Post-Keynesian economics*, ed. K.K. Kurihara. New Brunswick: Rutgers University Press.

Ricardo, D. 1817. On the principles of political economy and taxation. In *The works of David Ricardo*, ed. J.R. McCulloch. London: John Murray, 1846.

Ricardo, D. 1820. Funding system. In *The works and correspondence of David Ricardo*, vol. IV, ed. Piero Sraffa. Cambridge: Cambridge University Press, 1951.

Bureaucracy

Mancur Olson

Abstract

Bureaucracy in both businesses and governments continues to grow despite its unpopularity. Falling transport and communication costs have created global markets. The rising relative importance of firms with new technologies and methods often unsuited to market transfer via licensing of patents has given rise to multinational corporations with transnational bureaucracies. Government bureaucracies typically produce indivisible goods contributions to which by individual bureaucrats cannot be measured, giving rise to red tape and enabling bureaucracies to exploit society's demand for their products. Bureaucracies may not be highly efficient, but market failures that give rise to them also make them inevitable.

Keywords

Asymmetrical information; Bureaucracy; Bureaucratic growth; Business bureaucracy; Coase, R.; Communication costs; Economies of scale; Firm size; Indivisibilities; Market failure; Markets in the firm; Multinational firms; Niskanen, W.; Olson, M.; Output measurement; Patents; Portfolio investment; Profit centres; Red tape; Size of government; Social production function; Transaction costs; Transfer of technology; Transport costs; Tullock, G.; Weber, M.; Williamson, O.

JEL Classifications

H1

The study of bureaucracy has to deal with an elemental paradox. The role of bureaucracy has obviously increased dramatically in modern times. This is true not only of government bureaucracies but business bureaucracies as well. Though there were a few bureaucracies of significant size in pre-industrial times, such as the hierarchy of the Roman Catholic Church and the civil services of various Chinese empires, they were clearly exceptional. By contrast, a very large proportion of the total resources in the developed nations are controlled by either governmental or private bureaucracies. The role of governmental bureaucracies, at least, has increased with some rapidity within the last few decades. The increase in the use of bureaucracies has occurred in so many countries that it could hardly be due entirely to chance, and thus must be due to what are, in some sense, social choices to use more bureaucracy.

Normally, when there are great increases in the demand for or use of some product or instrumentality, this is accompanied by independent evidence of enthusiasm for the product or

instrumentality in question. When a society experiences a great increase in the demand for automobiles or for personal computers, there is at the same time a considerable amount of favourable commentary about whatever product is experiencing the boom in demand. There is pride in automobile ownership or awe at the power or compactness of personal computers. Nothing is more natural than that people's choices should be influenced by enthusiasms.

But where is the enthusiasm for bureaucracy that might have been expected to accompany the dramatic increase in the use of bureaucratic mechanisms? Any such enthusiasm is difficult to discern, and there are many conspicuous examples of dislike (or even contempt) of bureaucracy. Some of this negativism may be traced to particular ideological traditions, but this is not sufficient to explain the negativism; the problem is not only that the prevalence of the relevant ideology needs to be explained, but also that the lack of enthusiasm for bureaucracy prevails in a wide variety of ideological and cultural contexts and tends to apply (at least to some extent) to business as well as to governmental bureaucracies. There is no doubt that 'red tape' is viewed negatively by almost everyone, and that it is associated with bureaucracy, and especially governmental bureaucracy; the phrase is derived from the colour of the ribbons that were once used to tie folders of papers in the British government.

Some strands of the literature on bureaucracy are called into question by the paradox. Much of the admiring literature on bureaucracy is difficult to reconcile with the negative popular image of bureaucracy, whereas much of the negative literature suffers from the lack of any explanation of why virtually all societies, at least implicitly, keep choosing to use the instrumentality that is alleged to be so faulty.

Perhaps the most influential scholarly analysis of bureaucracy is not by an economist, but rather by the sociologist and historian, Max Weber. According to Weber:

> . . . the fully developed bureaucratic mechanism compares with other organizations exactly as does the machine with the non-mechanical modes of production . . .
>
> Precision, speed, unambiguity, knowledge of the files, continuity, discretion, unity, strict subordination, reduction of friction and of material and personal costs – these are raised to the optimum point in the strictly bureaucratic administration. (1946, p. 214)

Although also critical of 'bureaucratic domination', Weber's more positive view of bureaucracy has been influential in sociology and political science. Yet it does not appear to have generated systematic or quantitative empirical studies that have tended to provide any confirmation for it, and it surely is not in accord with the popular image of bureaucracy. Weber himself fails to identify any strong incentives in bureaucracies that would lead to efficient allocations of resources or to high levels of innovation.

Similarly, the popular pejorative view of bureaucracy is inadequate to the extent that it offers no explanation why modern societies choose or accept an increasing degree of bureaucratization. There is, admittedly, a rapidly growing economic literature on the growth of government that attempts to identify incentives that lead to a supra-optimal size of government. Examining this large literature would take us a long way from bureaucracy, and it has not in any case yet advanced to the point of generating a professional consensus on any incentive that would systematically bring about the overuse of government and thus of governmental bureaucracy, though some contributions (e.g. Mueller and Murrell 1985) are extremely promising. But even dramatic success in the literature on the growth of government would not be sufficient to solve the problem, as it would leave us with no explanation of the growth in modern times of business and other private bureaucracies.

Since an explanation of the growth of private bureaucracies is needed, and since an inquiry which begins with the growth of private bureaucracies may obtain some modest degree of detachment from the ideological controversies about the appropriate role of government, it may be best to consider private bureaucracies first. Here the basic question that must be answered is, 'Why do firms with hierarchies of employees exist?' Familiar economic theory explains that markets can under

the appropriate conditions allocate resources efficiently, so we must ask why individuals in the business hierarchy, and owners of the buildings and equipment that a typical corporation uses, do not use the price signals of the market to coordinate their everyday interaction. As Ronald Coase pointed out in somewhat different language in his seminal article on 'The Nature of the Firm' (1937), the survival of firms with hierarchies of long-run employees and long-term ownership of complementary fixed capital can only be explained by a kind of market failure. The type of market failure that Coase, and Williamson (1964, 1975, 1985) and the other economists that have developed the very important literature on private hierarchies have emphasized is 'transactions costs'. It would cost too much to contract out each day each of the very many separate tasks that are usually needed in any complex productive process, so in many cases it pays to forego the use of the market and to make long-term deals with employees who will perform such tasks each day as their superiors instruct them to do and receive in turn a regular salary. Though most of the literature in this tradition emphasizes only transactions costs, it is important to note that any market failure, such as that arising from an externality, could provide the incentive for the establishment of a firm that would internalize the externality, and all but the smallest firms have bureaucracies.

Though the foregoing argument also applies to small firms of the kind that predominated in preindustrial times, there have been some changes since the industrial revolution that, within this Coasian–Williamson framework, can provide important insights into the growth of business bureaucracies. One factor that made for larger and more bureaucratic firms was the discovery of technologies subject to indivisibilities that only a large enterprise can profitably exploit.

But the extraordinary improvement in the technologies of transportation and communication was probably far more important. Reductions in transportation and communication costs make it economic for firms to draw factors of production from farther away and also make it profitable for a firm to sell its output over a wider area. When transportation and communication technologies make it profitable for many firms to operate at a global rather than a village level, some very large firms can emerge. The improved transportation and communication also make it possible to coordinate the activities of a firm over a larger area. Superficial observers of the emergence of large firms have supposed that this growth of firm size entails a reduction in competition and a growth of monopoly. In fact, the dramatic reductions in transportation and communication costs have, of course, also increased the opportunities for market transactions over great distances, so the size of the market and the number of firms to which the typical consumer has access has (in the absence of extra trade barriers) also increased. At least in the Common Market or the United States, the average consumer, even if purchasing a product such as an automobile that is produced under greater-than-average economies of scale, has more firms competing for his business than did the average consumer in the typical rural village before the industrial revolution. Thus we see that the growth of business bureaucracy and the expansion of competitive markets are by no means necessarily obverse tendencies, but rather the kinds of things that often happen together.

The technologies that facilitated larger markets and larger firms also gradually led to the discovery of better methods of governing large-scale business organizations, as the historian Alfred D. Chandler has shown in some seminal historical studies of what he has called *The Visible Hand* (1977; see also 1962, 1980). Several of these innovations occurred in the unprecedently large and geographically scattered railroads in the 19th-century United States, and many involved the creation of separate 'profit centres' and other devices that enabled larger firms to use market mechanisms to fulfill some functions within the firm (Williamson 1985). This suggests that the costs and control losses in bureaucracies are still very considerable, so that business bureaucracy can only be explained in terms of rather substantial costs of using markets. The same conclusion emerges from the observation that activities that are highly space-intensive, such as most types of agricultural production, are quite resistant to

bureaucratization, even after the development of modern technologies of transportation and administration; the firms that succeed in surviving in most types of farming are normally too small to have bureaucracies (Olson 1985).

By contrast, in activities in which the transfer of new technologies and other information is especially important, market failure is likely to be fairly extensive, mainly because new information would only be rationally purchased by those who did not already have this information, and from this it follows that the market for new information is particularly handicapped by the asymmetrical information of the parties to any transaction. Thus, as J.C. McManus (1972), Buckley and Casson (1976) and, especially, Hennart (1982) have shown, the emergence of the multinational firms with bureaucracies that transcend national borders can be explained in this framework; capital can cross national borders through portfolio investment (almost all British and other foreign investment in the 19th century was portfolio investment), but the rise in the relative importance of firms with new technologies and methods that were often not well suited to market transfer via licensing of patents, gave rise to the multinational corporation.

The foregoing emphasis on the business bureaucracies that are generally neglected in discussions of bureaucracy makes possible a brief and unified explanation of governmental bureaucracy as well. Governmental bureaucracies are similarly necessary only because markets fail, at least to some degree; the theory of market failure is readily capable of being generalized to include all functions for which governmental are an efficient response (Olson 1986). Since governmental as well as market mechanisms are obviously imperfect, it does not follow from the presence of market failure that government intervention is normatively appropriate, since the government might fail even worse than the market, but market failures are nonetheless often important and always a necessary condition for optimal governmental intervention. Of course, it would be absurd to suppose that actual government intervention is always optimal or that governments always intervene when it is Pareto-efficient for them to do so. It is nonetheless instructive to look at the existence of government bureaucracy, as of business bureaucracy, in terms of market failure.

Among other reasons, it is instructive because the very conditions that give rise to market failure inevitably generate, in governments, and to a considerable degree also in firms, exactly those inefficiencies and rigidities that are popularly and correctly attributed to bureaucracies. Some of these inefficiencies also occur when either governmental or business bureaucracy is used inappropriately, but the problem is most easily evident, and most serious, in precisely those cases where market failure makes bureaucratic mechanisms indispensable.

The reasons why the same conditions that make markets fail also generate difficulties and inefficiencies in bureaucracies unfortunately do not lend themselves to brief exposition. But perhaps a faint and intuitive sense of the matter will be evident from a moment's reflection about what could make a bureaucracy necessary. If, say, the fruits or vegetables grown on a farm are best picked by hand and the best way to pay each worker is by the number of bushels picked, there is no need to have any bureaucratic mechanism for getting the work done. When piece-rate or commission systems of reward work well, the market gives each worker a more or less optimal incentive to work and to be as efficient as the worker knows how to be. In essence, the reason is that the output is highly divisible into more or less homogeneous units or the revenue attributable to each worker is known, and so the output of different workers can be measured with reasonable accuracy.

Let us now shift to an opposite extreme. Consider a typical civil servant in the foreign ministry of a government. Even supposing that the only purpose of the foreign ministry in question was peacefully to maintain the country's independence, there would still be a stupendous difficulty in rewarding the civil servant on a piece-rate or commission basis, or in any way that is proportional to his productivity. The security of the country in question would normally depend in large part on what might loosely be described as the state of the international system – on worldwide indivisible or public good for which no one

country could be entirely responsible. But even if the country in question were the only producer of this indivisible good, the foreign ministry would not be the only part of the government or the country that was relevant. Even in the foreign ministry, the typical civil servant is only one among thousands. How is his individual output to be measured, or even distinguished from that of his co-workers? The civil servant obviously cannot be paid in proportion to the revenue he generates, because if there really is market failure, the output cannot be sold in a market in the first place. Thus in practice, the remuneration of civil servants involved in producing public goods is not even a close approximation to each civil servant's true output; rewards in civil services will depend dramatically on proxy variables for performance such as seniority, education, and the fidelity of the employee to the interests of his superior and to the 'culture' or ideology of that bureaucracy. The peculiarities of civil service personnel systems, competitive bidding rules, and red tape are mainly explained by this logic (Olson 1973, 1974).

The knowledge of the 'social production function' of a government bureaucracy producing public goods will also be limited by the same indivisibility that has been described; there are fewer countries, or even airsheds for pollution abatement, than there are farms (or experimental plots at agricultural experiment stations), so in general less is known about how to run countries or control pollution than about agriculture or about production processes in other competitive industries (Olson 1982). The same indivisibility that obscures the social production function and the productivity of individual civil servants and other public inputs also insures that there cannot be even an imperfectly competitive market, so there is also no direct information on what an alternative bureaucracy could have achieved in the same circumstances.

In large part, it is the lack of information due to the indivisibilities described above that allows some of the bureaucratic pathologies described in Niskanen (1971) and Tullock (1965) to occur. In Niskanen's widely cited formal model, it is assumed that only the government bureaucrats know how many resources are required to produce a given public output. These bureaucrats are assumed to gain from growth of the bureaucracy, because an official's power, opportunities for promotion and other perquisites are assumed to be an increasing function of the budget the bureaucrat administers. An agency faces the constraint, however, that the electorate will not sustain any government programme whose total costs exceed the total value of its output. The optimization of government bureaucrats therefore leads to a bureaucracy far larger than is Pareto-efficient; in essence the bureaucracy takes all of the surplus under the society's demand curve for the government output at issue. Critics of Niskanen's model have pointed out that it neglects the subordination of bureaucrats to politicians, and that politicians whose opportunities for re-election are positively correlated with the government's performance will endeavour to prevent bureaucracies from taking all of the surplus (see, for example, Breton and Wintrobe 1975). These criticisms have substantial empirical support, but it is also true that there are many known cases where officials who fear a lower budget allocation than anticipated for their agency will eliminate or threaten to eliminate their politically most cherished activity rather than a marginal activity; this is precisely what Niskanen's model predicts. Though any final conclusion must await further research, the evidence available so far appears to suggest that the lack of information due to the indivisibilities described above does often allow bureaucracies to appropriate some of the surplus that consumers might otherwise be expected to receive, but that the incentives faced by politicians tends to keep bureaucracies from getting anything resembling the whole of this surplus.

Bureaucracies operating in a market environment share some of the information problems that confront government agencies providing public goods, but not others. The divisions of a large corporation that handle personnel, accounting, finance or public relations for the entire corporation provide collective goods to the corporation as a whole. They are in many ways in a situation analogous to the foreign ministry described above when deciding how much of the total profits of the firm to attribute to a given corporate employee;

this accounts for the many similarities of large corporate and civil-service bureaucracies. But the corporation as a whole, and even the nationalized firm producing private goods in a market, does not, when it sells its output, have as great a difficulty as the government agency that produces a collective of public output that is indivisible and unmarketable. The firm produces a good or service that is divisible in that it may be provided to purchasers and denied to non-purchasers. This means that the output is directly measurable in some physical units or at least that the revenue obtained from this output is measurable. Since consumers, even in the absence of any high degree of competition, will have alternative uses for their money, the private corporation or nationalized firm in a market economy will get some feedback about how much value it is providing. If there is no legal barrier to the operation of a competitive enterprise and the market is contestable, the society will also have at least potential information about what value an alternative organization could provide. An enterprise in the market produces an output from which non-purchasers may be excluded, and this also means there is normally better knowledge of the production functions for private goods than of production functions for public goods. All this implies that the problems of bureaucracy are less severe in private business than in government agencies producing public goods. Interestingly, they are also less severe in government enterprises that unnecessarily produce private goods that private firms would readily provide than they are in agencies that produce public goods that would not have been provided by the market. The more flexible personnel policies in some nationalized firms than in classical civil service contexts thus provides support for the conception offered here.

The paradox of a vast growth of both public and private bureaucracy at the same time that there is almost a consensus that bureaucracies are not very efficient or flexible, thus appears to have a resolution. There are fundamental reasons, arising from the inherent conditions causing market failure that make both public and private bureaucracies inevitable. These same reasons also explain why bureaucracies lack the information needed for high levels of efficiency. But these same market failures show that (though the existing degree of bureaucracy may of course be far from optimal), it should not be surprising that societies choose to use more private and public bureaucracy even as they condemn such bureaucracy.

See Also

- ▶ Hierarchy
- ▶ Public Choice
- ▶ Socialism
- ▶ Weber, Max (1864–1920)

Bibliography

Breton, A., and R. Wintrobe. 1975. The equilibrium size of a budget maximizing bureau. *Journal of Political Economy* 83: 195–207.

Buckley, P., and M. Casson. 1976. *The future of the multinational enterprise*. London: Macmillan.

Chandler, A.D. 1962. *Strategy and structure: Chapters in the history of American industrial enterprise*. Cambridge, MA: MIT Press.

Chandler, A.D. 1977. *The visible hand: The managerial revolution in American business*. Cambridge, MA: Harvard University Press.

Chandler, A.D., and H. Daems, eds. 1980. *Managerial hierarchies: Comparative perspectives on the rise of modern industrial enterprise*. Cambridge, MA: Harvard University Press.

Coase, R.H. 1937. The nature of the firm. *Economica* 4 : 386–405.N.S.

Hennart, J.-F. 1982. *A theory of the multinational enterprise*. Ann Arbor: University of Michigan Press.

McManus, J.C. 1972. The theory of the international firm. In *The multinational firm and the nation state*, ed. G. Paquet and D. Mills. Ontario: Collier Macmillan.

Mueller, D.C., and P. Murrell. 1985. Interest groups and the political economy of government size. In *Public expenditure and government growth*, ed. F. Forte and A. Peacock. Oxford: Basil Blackwell.

Niskanen, W.A. 1971. *Bureaucracy and representative government*. Chicago: Aldine-Antherton.

Olson, M.L. 1973. Evaluating performance in the public sector. In *The measurement of economic and social performance*, Studies in income and wealth, ed. M. Moss, vol. 38. New York: National Bureau of Economic Research, Columbia University Press.

Olson, M.L. 1974. The priority of public problems. In *The corporate society*, ed. R. Marris. London: Macmillan.

Olson, M.L. 1982. Environmental indivisibilities and information costs: Fanaticism, agnosticism, and

intellectual progress. *American Economic Review: Papers and Proceedings* 72: 262–266.
Olson, M.L. 1985. Space, agriculture, and organization. *American Journal of Agricultural Economics* 67: 928–937.
Olson, M.L. 1986. Toward a more general theory of governmental structure. *American Economic Review: Papers and Proceedings* 76: 120–125.
Tullock, G. 1965. *The politics of bureaucracy.* Washington, DC: Public Affairs Press.
Weber, M. 1946. Bureaucracy. In *From Max Weber: Essays in sociology*, ed. H. Gerth and C.W. Mills. New York: Oxford University Press.
Williamson, O.E. 1964. *The economics of discretionary behavior: Managerial objectives in a theory of the firm*. Englewood Cliffs: Prentice-Hall.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and anti-trust implications*. New York: The Free Press.
Williamson, O.E. 1985. *The economic institutions of capitalism*. New York: The Free Press.

Buridan, Jean (c1295–1356)

Martin Hollis

Abstract

French scientist and philosopher, Buridan studied philosophy with William of Ockham in Paris, where he became a professor and (in 1328 and 1340) rector. He made Ockham's nominalism the basis of an empirical physics, opposed to much in Aristotelian physics and paving the way for 17th-century mechanics. His *Consequentiae* addresses the theory of modal propositions in logic and attempts perhaps the first deductive derivation of the laws of deduction. But his name is best known from 'Buridan's Ass', the poor beast which is placed half-way between two identical bales of hay and starves to death for want of a reason for choosing one over the other. The example and name seem to have arisen later to refute his contention that will cannot operate without a sufficient reason, although Buridan himself mentions a dog in like state, in commenting on Aristotle's *De Coelo*. The topic is relevant to the theory of rational choice, where it could be awkward, if there were no way of resolving the problems of multiple equilibria. Presumably the ass must be allowed to have sufficient reason to pick at random, although that could still be awkward, if several asses had a coordination problem.

French scientist and philosopher, Buridan studied philosophy with William of Ockham in Paris, where he became a professor and (in 1328 and 1340) rector. He made Ockham's nominalism the basis of an empirical physics, opposed to much in Aristotelian physics and paving the way for 17th-century mechanics. His *Consequentiae* addresses the theory of modal propositions in logic and attempts perhaps the first deductive derivation of the laws of deduction. But his name is best known from 'Buridan's Ass', the poor beast which is placed half-way between two identical bales of hay and starves to death for want of a reason for choosing one over the other. The example and name seem to have arisen later to refute his contention that will cannot operate without a sufficient reason, although Buridan himself mentions a dog in like state, in commenting on Aristotle's *De Coelo*. The topic is relevant to the theory of rational choice, where it could be awkward, if there were no way of resolving the problems of multiple equilibria. Presumably the ass must be allowed to have sufficient reason to pick at random, although that could still be awkward, if several asses had a coordination problem.

Burke, Edmund (1729–1797)

C. B. Macpherson

Burke was born in Dublin and died at his estate at Beaconsfield. He is usually remembered as the champion of tradition, hierarchy, privilege and prejudice. His splendid defence of these in his

Reflections on the Revolution in France has so overshadowed all his other writings that a different side of his thought – his unqualified embrace of the capitalist market economy – has pretty well dropped out of sight. Yet his market orientation was clear enough to his contemporaries. Adam Smith is reported to have said of Burke 'that he was the only man who, without communication, thought on these topics exactly as he did'. The topics were the naturalness and beneficence of the capitalist market economy. That Burke the traditionalist and believer in Divine and Natural Law could praise the self-regulating capitalist market economy, and urge on the government a policy of laissez-faire, is at first sight incredible, so incompatible do the two positions appear to be. In fact, as we shall see, they are not incompatible. And certainly Burke saw no inconsistency, for he regarded the market economy as part of the natural order of the universe, and even as divinely ordained. In spite of his gibe in the *Reflections*, lumping together 'oeconomists' and 'sophisters', Burke was himself a skilled economist. Indeed, he boasted in a later work how much he had done 'in the way of political oeconomy', which he had made 'an object of my humble studies, from my very early youth to near the end of my service in parliament. . .'. He was fully aware that the capitalist economy, driven by (in his words) monied men's avarice, their desire of accumulation, their love of lucre, could be appallingly hard on the wage-labourer. Nevertheless, as he wrote in his *Thoughts and Details on Scarcity*, it would be 'pernicious to disturb the natural course of things, and to impede, in any degree, the great wheel of circulation which is turned by the strangely directed labour of these unhappy people. . .'. For 'labour is a commodity like every other, and rises or falls according to the demand. This is in the nature of things.' Hence 'labour must be subject to all the laws and principles of trade, and not to regulation foreign to them. . .'. We should not try to escape calamity by 'breaking the laws of commerce, which are the laws of nature, and consequently the laws of God. . .'. Burke's grasp of *political* economy becomes evident in his recognition that the laws of commerce could not operate unless the mass of the people was kept subordinate and unless they accepted that position as natural. Referring to the nation's need for continuous capital accumulation, he wrote:

> To be enabled to acquire, the people, without being servile, must be tractable and obedient. The magistrate must have his reverence, the laws their authority. The body of the people must not find the principles of natural subordination by art rooted out of their minds. They must respect that property of which they cannot partake. They labour to obtain what by labour can be obtained; and when they find, as they commonly do, the success disproportioned to the endeavour, they must be taught their consolation in the final proportions of eternal justice. Of this consolation whoever deprives them, deadens their industry, and strikes at the root of all acquisition as of all conservation. He that does this is the cruel oppressor. . ..

It was Burke's genius to see that the operation of the capitalist market economy *required* that the hard core of a hierarchical order should be maintained, and should be accepted even by those it oppressed. And the need for a hierarchical order could most easily be put as a need for the traditional order. So there is no inconsistency between his praise of tradition and his praise of the market economy.

References

Kramnick, I. 1977. *The Rage of Edmund Burke: Portrait of an ambivalent conservative*. New York: Basic Books.

Macpherson, C.B. 1980. *Burke*. Oxford: Oxford University Press.

Of Burke's own works the most famous is the *Reflections on the Revolution in France* (1790); a modern reprint, in Pelican Classics, has an introduction by Conor Cruise O'Brien. Burke's equally important *Thoughts and Details on Scarcity* (1795) is only available in scarce editions of his Collected Works. Adam Smith's remark is reported in Robert Bisset, *Life of Edmund Burke,* 2nd edn, London 1800, vol. 2, p. 429.

O'Gorman, F. 1973. *Edmund Burke, his political philosophy*. London: Allen & Unwin.

Three recent studies of Burke's thought are:

Burns, Arthur Frank (1904–1987)

Geoffrey H. Moore

Keywords

Burns, A. F.; Business cycle measurement; Business cycles; Council of Economic Advisers (USA); Economic Policy Advisory Board (USA); Federal Reserve System; Mitchell, W. C.; National Bureau of Economic Research

JEL Classifications

B31

Burns was born in Stanislau, Austria, on 27 April 1904. In 1914 his family emigrated to the United States, settling in Bayonne, New Jersey. Burns became a member of the economics faculty at Rutgers University in 1927, leaving in 1941 to accept an appointment at Columbia University, where he taught for many years and became John Bates Clark Professor of Economics Emeritus. He joined the staff of the National Bureau of Economic Research in New York in 1930, was director of research, 1945–53, and president 1957–67. In Washington Burns served as chairman of the Council of Economic Advisers, 1953–6; Counsellor to the President, 1969–70; chairman of the Federal Reserve System, 1970–78; and member of the President's Economic Policy Advisory Board since 1981. From 1981 to 1985 he was US Ambassador to the Federal Republic of Germany. In 1978–80 and again after 1985 he was distinguished scholar in residence at the American Enterprise Institute.

Burns's economic studies have been primarily concerned with economic growth, business cycles, inflation, and economic policies bearing upon these phenomena. In *Production Trends in the United States since 1870,* published in 1934, he examined growth rates in individual industries, noting the nearly universal tendency towards retardation. An initial stage of rapid growth in a new industry is usually followed by slower growth as it loses part of its market or its resources to still newer industries. Despite the tendency towards slower growth and eventual decline of most industries, Burns noted that this did not imply that growth in total output would slow. The underlying cause, that is the rise of new industries, would itself help to maintain rapid growth in total output.

Burns's collaboration with Wesley Mitchell in the study of business cycles led to many innovations in measurement technique and to a vast accumulation of knowledge about the characteristics of cycles and the economic interactions that generated them. It also led to a more realistic view of what business cycle theory had to explain and what economic policy could be expected to accomplish. This in turn was useful to Burns in his later role as an economic policymaker, that is as a presidential adviser and as chairman of the Federal Reserve. Before taking on these responsibilities he wrote prophetically (1953): 'It is reasonable to expect that contracyclical policy will moderate the amplitude and abbreviate the duration of business contractions in the future . . . But there are no adequate grounds, as yet, for believing that business cycles will soon disappear, or that the government will resist inflation with as much tenacity as depression . . .' Burns's subsequent efforts were largely directed to improving the anti-recession, anti-inflation, and growth promoting policies of government.

Selected Works

1930. *Stock market cycle research*. New York: Twentieth Century Fund.

1934. *Production trends in the United States since 1870*. New York: National Bureau of Economic Research.

1938. (With W.C. Mitchell.) *Statistical indicators of cyclical revivals*. New York: National Bureau of Economic Research, Bulletin 69.

1946. (With W.C. Mitchell.) *Measuring business cycles*. New York: Columbia University Press.

1946. *Economic research and the Keynesian thinking of our times*. New York: National Bureau of Economic Research, 26th Annual Report.

1950. *New facts on business cycles*. New York: National Bureau of Economic Research, 30th Annual Report. Reprinted in Burns (1969).

1952. *The instability of consumer spending*. New York: National Bureau of Economic Research, 32nd Annual Report.
1952. (et al.) *Wesley Clair Mitchell: The economic scientist*. New York: Columbia University Press.
1953. *Business cycle research and the needs of our times*. New York: National Bureau of Economic Research, 33rd Annual Report.
1954. *The frontiers of economic knowledge*. Princeton: Princeton University Press.
1957. *Prosperity without inflation*. New York: Fordham University Press.
1960. Progress towards economic stability. *American Economic Review* 50: 1–19.
1966. *The management of prosperity*. New York: Columbia University Press.
1967. (With P.A. Samuelson.) *Full employment, guideposts, and economic stability*. Washington, DC: American Enterprise Institute.
1968. (With J.K. Javits and C.J. Hitch.) *The defense sector and the American economy*. New York: New York University Press.
1968. Business cycles. *International encyclopaedia of the social sciences*. New York: Macmillan and Free Press. Reprinted in Bums (1969).
1969. *The business cycle in a changing world*. New York: Columbia University Press.
1978. *Reflections of an economic policy maker*. Washington, DC: American Enterprise Institute.
1984. The American trade deficit in perspective. *Foreign Affairs* 62(5):1058–1069.
1985. Interview: an economist's perspective over 60 years. *Challenge* 17–25.

Burns, Arthur Robert (1895–1981)

Henry W. Spiegel

Keywords
Burns, Arthur R.; Burns, Eveline M.; Competition theory

A native of London, Burns was educated at the London School of Economics, where he was a pupil of Edwin Cannan. His doctoral dissertation, *Money and Monetary Policy in Early Times*, appeared in a prestigious series in 1927 and is still a standard work on the subject. After the completion of his studies Burns moved to the United States and taught at Columbia University from 1928 to 1963. His service there overlapped with that of his wife Eveline M. Burns, with whom he published an introductory economics text in 1928, and with that of Arthur F. Burns, another noted economist.

In 1936 Burns, still an assistant professor, published *The Decline of Competition*, the bulk of which consisted of chapters on trade associations, price leadership, market sharing, price stabilization, price discrimination, non-price competition and integration. The work formed part of a discussion that had been set in motion by the writings of Sraffa, Joan Robinson and Chamberlin and which explored the no man's-land between competition and monopoly. It was to serve as a bridge that linked the abstractions of the theories of imperfect or monopolistic competition with the world of reality. Standing between abstraction and description, Burns' work was in the main an attempt at classification. It holds middle ground between the soaring abstractions of pure theory and the industry studies published by Walton Hamilton and Associates under the title *Price and Price Policies* in 1938. Hamilton was a follower of Veblen. Burns shared a friendly disposition toward institutional economics with other Columbia economists.

The Decline of Competition constitutes Burns's main claim to fame. In later years he directed a Twentieth Century Fund study of electric power and government policy, and in 1955 he published *Comparative Economic Organization*. The former work was overtaken by the rise of atomic power as a source of electric energy, and the latter compared in isolation various factors affecting the national income.

Selected Works

1927. *Money and monetary policy in early times*. London: Kegan Paul; New York: Knopf.

1928. (With Eveline M. Burns). *Economic world*: *A survey*. London: Oxford University Press.
1936. *The decline of competition*. New York: McGraw-Hill.
1948. (Director of Research). *Electric power and government policy*. New York: Twentieth Century Fund.
1955. *Comparative economic organization*. New York: Prentice-Hall.

Bibliography

Hamilton, Walton H. and Associates. 1938. *Price and price policies*. New York: McGraw Hill.

Burns, Emile (1889–1972)

Sam Aaronovitch

Burns was born in St Kitts, where his father was in the colonial administration. He was educated in London and Cambridge and earned his living in the shipping industry before working full-time for the Labour Research Department (1925–9) and later as an official of the Communist Party of Great Britain until he retired. He was a gifted expositor and popularizer of economic issues and problems and the books and pamphlets he wrote when at the Labour Research Department were much used in the trade union movement. They included a study of the textile industry (requested by the United Textile Workers Union), syllabuses on *Finance* and on *Imperialism* and a book on *Modern Finance* (1920) which set out the workings of the financial system.

He was active in the Independent Labour Party and then joined the Communist Party in 1921, an action which reinforced his interest in the contribution of Marxist economics to understanding contemporary economic problems. During the General Strike of 1926 he became propaganda organizer of the St Pancras Council of Action and published an invaluable survey of the work of Trades Councils and Councils of Action during the strike (1926).

He did not produce any substantial work of economic analysis but responded to or commented on key economic events and debates with a labour movement audience very much in his sights. *The Crisis: The Only Way Out* (1932) dealt forcibly with several of the favoured nostrums of the day such as 'managed credit', 'controlled and planned capitalism', the Fordist 'high wage' proposal etc., but in following the line of the CPGB at the time could only put forward as the solution a socialist revolution and fully planned economy on the Soviet model.

In 1940 he wrote a reply to Keynes's pamphlet *How to Pay for the War* (*Mr Keynes Answered*, 1940), which he saw as primarily a method of cutting workers' real wages in an imperialist war and as preparation for the slump which Emile Burns believed (and thought the capitalists also believed) would certainly and swiftly follow the end of the war. This was a widely held view among economists at the time. He treated the notion of 'equal sacrifice' in such an unequal society with the scorn it deserved.

Continuing his keen interest in financial and monetary issues his last book, published in 1968, was entitled *Money and Inflation*; he attacked both cost push and demand led theories of inflation, focusing rather on the *interests* of capital in using price increases as a way of cutting real wages, as offering a more plausible explanation.

His gift of clear exposition was shown in his *Introduction to Marxism* (1952), which went through many editions.

Whatever the direct impact of his writings on contemporary problems it is likely that his most lasting influence in the field of ideas came from his role as an editor and translator of the work of Marx and Engels. His compilation *Handbook of Marxism* published in the 1930s by the Left Book Club was widely circulated and contributed much

to the spread of Marxist ideas in that period. Of special interest to economists was his translation of Part One of Marx's *Theories of Surplus Value* (sometimes described as Part IV of *Capital*), published in 1964 by the Foreign Languages Publishing House in Moscow and issued in Britain by Lawrence and Wishart.

His work as an economist was clearly limited by his continuous absorption in political and administrative activity (at one period he edited the busmen's rank and file journal) and his intense desire to serve all those struggling for improved conditions and deeper understanding of their problems.

Selected Works

1920. *Modern finance,* 2nd ed., revised. London: Oxford University Press, 1922.
1922. *Finance: An introductory course for classes and study circles*. London: Labour Research Department.
1926. *The general strike, May 1926: Trades councils in action*. London: Labour Research Department.
1927. *Imperialism. An outline course*. London: Labour Research Department.
1932. *The crisis: The only way out*. London: Martin Lawrence.
1935. *A Handbook of Marxism: Being a collection of extracts from the writings of Marx, Engels, and the greatest of their followers*. London: Victor Gollancz.
1939. *What is Marxism?* London: Victor Gollancz. Revised edn as *An introduction to Marxism.* London: Lawrence and Wishart, 1952.
1940. *Mr Keynes answered: An examination of the Keynes plan*. London: Lawrence & Wishart.
1964. *Theories of surplus value: Volume IV of capital* (trans: Marx, K.). Moscow/London: Foreign Languages Publishing House/Lawrence & Wishart.
1968. *Money and inflation*. London: Lawrence & Wishart.

Burns, Eveline Mabel (1900–1985)

W. Cohen

Eveline M. Burns was Professor of Social Work at the School of Social Work of Columbia University (1946–67) and Lecturer in the Department of Economics (1928–42). She received her Ph.D. in economics from the London School of Economics (1926) and was elected an Honorary Fellow of the School (1967). She was consultant to the Employment Opportunities Staff of the Committee on Economic Security (1934–5) and Chief of the Economic Security and Health Section of the National Resources Planning Board (1942).

Burns performed a significant role in translating the complexities of social security issues and programmes to both economists and the social welfare community and in influencing the development of income security policy in the United States. She was a member of several governmental advisory and research agencies and an adviser to numerous social welfare groups. She lectured on and published some of the early books and articles on social security and thus had an important impact on the administrators and future leaders during the initial period when the provisions of the Social Security Act were being implemented.

Selected Works

1936. *Towards social security*. New York and London: McGraw-Hill Book Company.
1941. *British unemployment programs 1920–1938*. Washington, DC: Social Science Research Council.
1942. *Security, work and relief problems*. A report issued by the National Resources Planning Board. Washington, DC: US Government Printing Office.
1949. *The American social security systems*, 2nd ed. Boston: Houghton-Mifflin, 1951.

1956. *Social security and public policy.* New York: McGraw-Hill Book Company.
1973. Health services for tomorrow: trends and issues. New York: Dunellen.

Business Cycle Measurement

Don Harding and Adrian Pagan

Keywords

Burns, A.; Business cycle; Business cycle measurement; Censoring operations; Coincident indices; Crossing points; Data filters; Fluctuations vs cycles; Growth cycles; Markov switching (MS) processes; Mitchell, W.; Periodic cycles; Random variables; Reference cycle; Spectral analysis; Turning points

JEL Classifications

E32

Measurement of business cycles provides a reference point against which macroeconomic theories and policy discussion can be assessed. The process requires an operational definition of a cycle, criteria to distinguish business cycles from other forms of fluctuation, procedures to detect the presence of a business cycle, and methods to measure its features. A central theme of this entry is that good measurement should not prejudge the nature of the phenomena under investigation. Moreover, it should produce statistics which are informative about features of interest and which can be formally analysed.

Defining and Detecting Cycles

In their classic work *Measuring Business Cycles*, Burns and Mitchell (BM) (1946) define *specific cycles* in a series y_t in terms of *turning points* in its sample path. This tradition has been central to work at the NBER and other institutions such as the IMF (2002) and the OECD (leading indicators). When it came to discussing the business cycle, BM simply referred to y_t as the *level* of aggregate economic activity, although in this article we will regard it as the *log* of economic activity, as the turning points in the level and the log of economic activity are the same. When Mintz (1969, 1972) had trouble finding turning points in the level of activity in surging economies such as West Germany's, this led her to first extract a permanent component p_t from y_t and to then study turning points in $z_t = y_t - p_t$. The resulting *growth cycle* in z_t has many forms depending on the method used to extract the permanent component. Others, such as the Economic Cycle Research Institute (ECRI) (growth rate cycle), have studied turning points in the differenced data $\Delta^r y_t$.

A generalization of this, explored by Kedem (1980, 1994) and Harding (2003), is to study turning points in $\Delta_r y_t$.

At the time Mitchell began his work, the alternative way of thinking about cycles (or oscillations) was to view y_t as composed of periodic components represented by sine and cosine waves, that is

$$y_t = \sum_{j=1}^{m} \alpha_j \cos \lambda_j t + \beta_j \sin \lambda_j t, \qquad (1)$$

where λ_j is the frequency of the *j'th* oscillation. If $m=1$ there would be a single periodic cycle. The problem with this way of looking at cycles was that few economic time series showed evidence of periodicity. To overcome that problem α_j and β_j were allowed to vary stochastically over time. Specifically, they were treated as uncorrelated random variables with zero mean and variance σ_j^2. This formulation meant that y_t had to be a stationary random variable and so could not be applied to the levels of variables such as GDP (unlike turning point analysis). However, in this form one can measure the importance of the j^{th} periodic cycle by looking at the ratio of σ_j^2 to the variance of y_t and it is the basis of spectral analysis. Such a perspective has increasingly been referred to as studying fluctuations rather than

cycles, since the focus of attention is upon the variance of y_t.

To understand the difference between these alternative ways of measuring cycles, take the special case where $\lambda_1 = 0$ and there is another frequency λ_2. Then

$$\begin{aligned} y_t &= \alpha_2 \cos \lambda_2 t + \beta_2 \sin \lambda_2 t + \alpha_{1t}, \\ &= y_t' + \alpha_{1t} \end{aligned} \tag{2}$$

Now there are certainly turning points in the series y_t' and the period between them is determined by λ_2. In contrast, the turning points in y_t will also be affected by the random variable α_{1t}, and thus may be very different to those in y_t'. Information about cycles gathered from spectral analysis concerns the nature of turning points in y_t' and not y_t. To give a more concrete illustration of this point, suppose that the model for y_t is of the form

$$y_t = 1.4y_{t-1} - .53y_{t-2} + e_t.$$

Then the periodic cycle in y_t can be isolated by setting $e_t = 0$ to get y_t'. To use the dating methods of an institution like NBER, the turning points in y_t' are 22 quarters apart, as could also be discovered by computing the roots of $(1 - 1.4L + .53L^2) = 0$. However, applying the same methods to y_t, one finds that the turning points in y_t will be on average 12 quarters apart. A further disadvantage of the periodic cycle approach is that the data needs to be filtered to render it stationary before analysis proceeds and, as Cogley observes elsewhere in this dictionary (data filters), the filters most commonly used by macroeconomists can introduce spurious periodic cycles, thereby blurring the picture.

Locating Turning Points

To locate turning points in a series it is necessary to define what these are and to provide some way of recognizing them in a given data-set. An obvious solution is to use the idea that peaks (troughs) are local maxima (minima) in the series y_t. Hence, if $\vee_t(\wedge_t)$ are binary variables taking the value of unity where there is a peak (trough) at t and zero otherwise, applying the proposed definition gives

$$\vee_t = 1\left(y_t < y_{t\pm j},\ 1 \le j \le k\right) \tag{3}$$

$$\wedge_t = 1\left(y_t > y_{tj\pm},\ 1 \le j \le k\right) \tag{4}$$

In Eqs. (3) and (4) $1(A)$ is the indicator function taking the value 1 if the event A is true and zero otherwise. Of course, this still leaves one with the need to describe the interval over which the local maxima or minima are said to occur, that is, a choice needs to be made regarding k. To replicate the main features of Burns and Mitchell's specific cycle dating procedures, it is necessary to set $k = 5$ for monthly data or $k = 2$ for quarterly data.

This is not the last of the choices that need to be made when locating turning points, but the others do not relate to the location of local maxima and minima. Rather, they concern the question of whether one should eliminate some of the local turns in deciding on a final set of turning points. Mostly these extra restrictions are imposed as phase length constraints, where phases are the periods of expansions and contractions between turning points. Thus, NBER dating procedures require that completed phase and complete cycles durations last longer than 5 and 15 months respectively. These are generally referred to as censoring operations. Whether turning points should be censored depends on the objectives of the research. If the objective is to match NBER business cycle dates, then censoring is essential. But if the researcher is pursuing other objectives such censoring may not be necessary. Censoring turning points makes it much harder to formally analyse the statistics produced and this may provide an important reason for not imposing them.

BM acknowledged that the final set of dates they selected for turning points reflected considerable amounts of judgement and incorporated specific information about economic activity at particular dates. Today, academic economists are primarily interested in the average characteristics of the cycle, and so it may well be that automated

methods of turning point detection become attractive. In the early post-Second World War period many of the procedures used by BM were codified, producing an expert system for locating turning points. Ultimately, Bry and Boschan (1971) produced an algorithm and FORTRAN program (called BB here) that largely replicated this expert system. Subsequently Mark Watson (1994) implemented this algorithm in the language GAUSS, and that code is available at (http://www.princeton.edu/mwatson).

There were three key components to the BB algorithm. The first was to engage in some smoothing of the series and to find an initial set of turning points using Eqs. (3) and (4) with $k = 5$. The second was to eliminate enough of these turning points so as to ensure that expansion and contraction phases exceeded 5 months in duration, while completed cycles exceed 15 months in duration. The third component was to ensure that peaks and troughs alternated by deleting multiple sequential occurrences of these. That was done through the application of various rules, such as choosing between two peaks based on which had the higher value of y_t.

Although BB were interested in analysing monthly data, they suggested a method for working with quarterly data that involved treating the observations on each of the months in a quarter as one-third of the quarterly value. A variant of BB has been developed by Harding and Pagan (2002) and called BBQ. It omits the smoothing in the BB algorithm but retains the three key principles of the BB algorithm. It also sets $k = 2$ and makes the minimum phase and cycle lengths two and five quarters respectively. Faster recursive algorithms for locating turning points have been developed by Artis et al. (2004) and James Engel. Engel's computer programs are called MBBQ. They are written in MATLAB and GAUSS and are available at the National Centre for Econometric Research (MBBQ Code).

Model-Based Procedures for Defining and Locating Turning Points

The procedures above do not require any knowledge of the data-generating process for y_t. An alternative approach is to adopt a model of Δy_t and use this to locate turning points. To date the models used are *parametric* and generally feature two regimes. Perhaps the best known parametric model is that of Hamilton (1989), where the growth rate is treated as a Markov switching (MS) process of the form $\Delta y_t = \mu_0(1 - \xi_t) + \mu_1 \xi_t + e_t$. Here μ_j are the growth rates in the two regimes, and these are indexed by a latent binary state, ξ_t, while e_t is a normally distributed zero mean error term. Here μ_0 is the growth rate of the low growth state and μ_1 is the high growth rate. Sometimes the restriction $\mu_0 < 0$ is also imposed. The model is completed by specifying the transition probabilities of moving from $\xi_{t-1} = 0$ or 1 to $\xi_t = 1$ or 0. The model can be made more complex with extra dynamics, different variances in each regime, allowing the transition probabilities to depend on some observable data, and so on. This parametric model is used to compute the conditional probability, $\Pr[\xi_t = 1 | A_t]$, where A_t is either all or a subset of the growth rates $\left\{\Delta y_j\right\}_{j=1}^{T}$. Thus the estimate of $\Pr[\xi_t = 1 | A_t]$ is a function of whatever growth rates are in A_t. Generally this probability will be a nonlinear function of the elements in A_t although a linear function can be quite a good approximation – see Harding and Pagan (2003) for an example.

The cycle is then associated with a binary variable S_t that takes the value 1 in expansion and zero in contraction. A rule is used to construct S_t by comparing the estimated probability of being in the high growth state with some critical value. Hamilton chose .5, and most of those using the technique have followed suit. Consequently, if $\Pr[\xi_t = 1 | A_t] > .5$, an expansion is signified and S_t is set to unity. If the criterion is not satisfied S_t is set to zero. Notice that the ξ_t are not the phase states; the latter are S_t. They are simply a device for producing some nonlinear structure in Δy_t, although often one can think of the outcomes for ξ_t as signifying a low or high growth period. The correlation between S_t and ξ_t may be very low. Many applications of this methodology have now been made and the MS model that one chooses seems to vary a lot with the series it is being applied to. The simple one described above rarely works satisfactorily.

In most instances a decision about the utility of the method is made by comparing the business cycle states produced by the rule based on the magnitude of $\Pr[z_t = 1 | A_t] > .5$ with those found by turning point methods. Because of the latter comparison one has to ask what the advantages there are in using a model to locate turning points. Chauvet and Piger (2003) claim that an advantage of the model-based approach is that it allows an investigator to forecast turning points in real time. There is some truth to this but it is exaggerated. Since forecasts can be found for any such model, they could be passed through any chosen dating algorithm to determine the predicted phases.

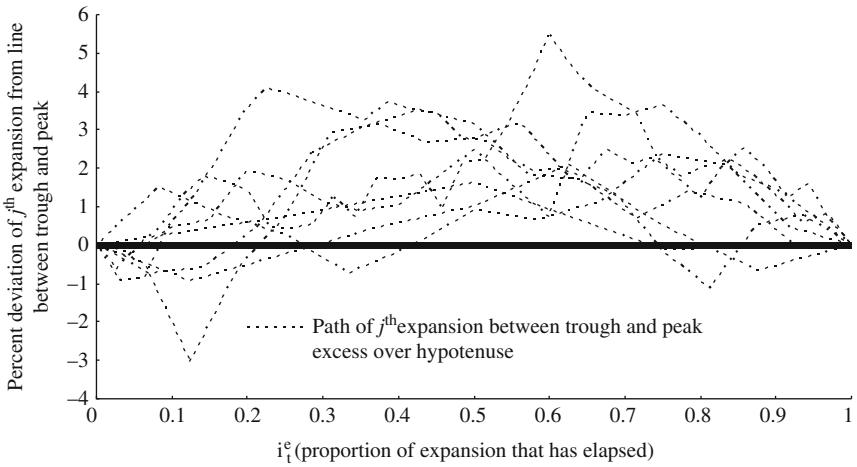
Measuring Cycle Features

Turning points segment time series into phases. An expansion phase runs from the trough to the next peak. A contraction runs from a peak to the next trough. In what follows it is easiest to just describe the derivation of information on expansions.

The two most basic statistics related to phases are duration and amplitude. The *duration* of an expansion is the number of periods of time between the trough and next peak. The *amplitude* of an expansion measures the change in y_t from trough to the next peak. In many cases y_t is the log of some variable such as GDP or industrial production, that is, $y_t = \ln(Y_t)$, and the amplitude has a natural interpretation as the approximate percentage change in Y_t between trough and peak.

Duration and amplitude form two sides of a triangle. Connecting the trough and peak produces the hypotenuse. If $y_t = \ln(Y_t)$, then the hypotenuse represents the path followed by a variable that exhibits a constant growth rate during an expansion. With this in mind it is instructive to inspect the actual path followed by the data, and to compare that path with the constant growth path represented by the hypotenuse. Figure 1 shows how US expansion paths have deviated from the constant growth rate path in the post-Second World War period. The important feature evident in this figure is that the growth rate of GDP is not constant over the expansion phase and typically is highest in the first half of an expansion.

While comparisons such as that in Fig. 1 are visually informative, there is also a need for statistics that summarize the average shape of phases. Sichel (1994) divides expansions into three stages, computes the average growth rate for each stage, and shows graphs of these, as well as providing formal statistical tests of equality of the growth rates in each stage. Harding and Pagan (2002) compare the cumulated gain in an expansion with what it would have



Business Cycle Measurement, Fig. 1 Deviation of sample path from hypotenuse: US GDP during expansions in the post-Second World War period (*Source*: Harding (2003))

been if growth had been constant throughout the phase. This comparison was motivated by the idea mentioned above, that a plot of y_t against t during an expansion would look like a triangle if growth had been constant. The area of such a triangle would be one-half the product of the amplitude and duration. If growth was not constant the area under the path actually followed by activity during the expansion would differ from the triangle. Thus, a comparison of the two areas provides a measure of the extent of departure from a constant growth scenario. The evidence seems to be that expansions do not feature constant growth in some countries like Australia, the United States and the UK, but do so in many Continental European countries. The shape analysis is interesting since a linear process for Δy_t will produce phases that, on average, have constant growth rates. So a failure to see this signals the need for a nonlinear process for Δy_t. The shape analysis also provides a useful tool for testing whether nonlinear models produce realistic business cycles.

All of the methods for summarizing business cycle information can be applied to growth cycles and to data that have undergone higher-order differencing. In addition, Sichel (1993) suggested tests for 'deepness' and 'steepness' in the growth cycle that were effectively tests for symmetry in the densities of z_t and Δz_t.

Using Multivariate Information in Defining and Detecting Business Cycles

Burns and Mitchell's famous definition of a business cycle – 'Business cycles are a type of fluctuation found in the aggregate economic activity of nations. . .a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general . . . contractions. . .' (1946, p. 1) – has two aspects. One points to the need to identify aggregate economic activity, and the other to the fact that there should be synchronization across many series during the phases of a business cycle. Burns and Mitchell commented that GDP was a suitable index of economic activity, although others, such as Moore and Zanrowitz (1986), have preferred a weighted average of several series rather than a single one. However, since data on GDP was not available to Burns and Mitchell, for either the time period or the frequency in which they were interested, it is natural that they placed more emphasis upon the second component of their definition when discussing the business cycle.

This second component emphasizes synchronization of the cycles in the specific series taken to represent economic activity. Burns and Mitchell took the turning points in many series and then extracted a *reference cycle* by determining those dates which peaks and troughs 'clustered around'. So a primary task is to be able to measure the tightness of the clusters. At the end of the process one also wishes to know how synchronized each of the specific cycles is with the cycle in the aggregate.

Harding and Pagan (2006) develop procedures to measure the tightness of clusters of turning points and the degree of synchronization of cycles through concordance indices that measure the fraction of time spent in the same phase. They apply those procedures to the series referred to by the NBER when dating the business cycle, and find that the turning points in those series are tightly clustered together. Harding (2003) finds that between March 1949 and September 2001 there is a concordance of 0.96 between the NBER business cycle states and the cycle obtained by locating turning points in US GDP.

Automated Construction of the Reference Cycle

To automate the calculation of the reference cycle requires some rules which will distill the specific cycle turning points into a single set of turning points. To determine what these rules might be, one could look at the NBER Business Cycle Dating Committee procedure. It has a similar modus operandi to that of Burns and Mitchell, as seen in

its discussion about dating the 2001 recession (NBER 2003). However, one rarely gets a precise description either of how its decisions are made or of the series used in that process. In addition, it seems as if the series which have been most influential in decisions may have been different at different periods in time. The clearest description of the procedures for aggregating turning points in a set of series to create a reference cycle is in Boehm and Moore (1984), who explain how NBER methods were used when establishing a reference cycle for Australia. Their description can be taken as authoritative because Moore was a pivotal figure in the NBER Business Cycle Dating Committee for many years. Moore and Zanrowitz (1986) also provide information on methods used by NBER in dating the business cycle.

Given that the process for establishing the reference cycle is a little vague, it should not be surprising that there have been few attempts at producing automated dating algorithms to establish it from multivariate series. Harding and Pagan (2006) construct an algorithm to replicate the NBER procedures described by Boehm and Moore (1984). They obtain the 'clustering parameter' which is essential to measuring the tightness of turning point clusters by looking at Boehm and Moore's spreadsheets. The resulting algorithm has produced a reference cycle that matches the Australian version established by Boehm and Moore quite well. Subsequently, it has been tested on US data, and is able to produce quite a good replication of the reference cycle for the United States, even though the clustering parameter had been calibrated with Australian data.

Model-Based Procedures for Defining Detecting and Extracting a Reference Cycle

Recently, academic economists have used parametric models to construct a coincident index and the reference cycle from n multivariate series $\Delta y_{1t}, \ldots, \Delta y_{nt}$. A common element to all approaches is to write Δy_{jt} as a function of a common component Δf_t and idiosyncratic components $u_{jt} (j = 1, \ldots, n)$. Hence a simple representation would be $\Delta y_{jt} = a_j \Delta f_t + u_{jt}$. The f_t is often thought of as the coincident index of the business cycle. Of course, there may be more than one f_t but, ultimately, we can think of combining them to form a single variable. There are then many ways that models for Δf_t and u_{jt} might be specified, depending upon how strong the assumptions are that one wishes to make about the nature of f_t and u_{jt}. Often Δf_t is given an MS form (for example, Chauvet and Piger 2003). Depending on what these assumptions are, they will determine how an estimate of f_t is to be made. Stock and Watson (1991) and Chauvet (1998) represent different approaches. In some instances one can avoid specifying precise parametric models for f_t and u_{jt}, restricting them only to be in a general class. Forni et al. (2001)'s dynamic factor approach is the main representative of this latter technique. The main issue with these approaches is that the coincident index and reference cycle obtained are conditioned on the assumptions made about the data-generating process. For that reason these approaches cannot provide a neutral measurement of the reference cycle.

Conclusion

Although widely used in official circles, Burns and Mitchell's methods of measuring cycles through turning points have been less popular in academia. But this has changed in recent years. There are a number of reasons why the methods have become increasingly attractive. First, information about the nature of the cycle phases can be generated, and this shape information proves important when one tries to construct models of economic activity. Second, the literature now contains expert systems for locating turning points, and these have been coded into various computer languages, thereby eliminating the judgmental aspect of the method. Nevertheless, the automatically generated turning points have been quite good approximations to those found via

judgment. Third, the ability to produce simulated data from parametric models means that such information can be passed through the algorithms for locating turning points to produce simulated distributions for the statistics that summarize the features of the cycle. Fourth, the emerging mathematics literature on crossing points provides a natural foundation on which to build a distribution theory for Burns and Mitchell's methods. Fifth, there is now a large literature on parametric methods for locating turning points and measuring cycles. This latter literature can readily be linked to the nonparametric turning point approach of investigators such as Burns and Mitchell, as seen in Harding and Pagan (2003).

See Also

- ▶ Burns, Arthur Frank (1904–1987)
- ▶ Data Filters
- ▶ International Real Business Cycles
- ▶ Mitchell, Wesley Clair (1874–1948)

Bibliography

Artis, M., M. Marcellino, and T. Proietti. 2004. Dating the Euro area business cycle: A methodological contribution with an application to the Euro area. *Oxford Bulletin of Economics and Statistics* 66: 537–565.

Boehm, E. and Moore, G. 1984. New economic indicators for Australia, 1949–84. *Australian Economic Review* 4th quarter, 34–56.

Bry, G., and C. Boschan. 1971. *Cyclical analysis of time series: selected procedures and computer programs. Technical Paper No. 20*. New York: NBER.

Burns, A., and W. Mitchell. 1946. *Measuring business cycles*. New York: National Bureau of Economic Research.

Chauvet, M. 1998. An econometric characterization of business cycle dynamics with factor structure and regime switches. *International Economic Review* 39: 969–996.

Chauvet, M., and J. Piger. 2003. Identifying business cycle turning points in real time. *Federal Reserve Bank of St. Louis Review* 85(2): 47–61.

ECRI (Economic Cycle Research Institute). Growth rate cycle. Online. Available at http://www.businesscycle.com. Accessed 8 Sept 2006.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. 2001. Coincident and leading indicators for the EURO area. *Economic Journal* 101: 62–85.

Hamilton, J. 1989. A new approach to the economic analysis of non stationary time series and the business cycle. *Econometrica* 57: 357–384.

Harding, D. 2003. Essays on the business cycle. Ph.D. thesis, Yale University.

Harding, D., and A. Pagan. 2002. Dissecting the cycle: A methodological investigation. *Journal of Monetary Economics* 49: 365–381.

Harding, D., and A. Pagan. 2003. A comparison of two business cycle dating methods. *Journal of Economic Dynamics and Control* 27: 1681–1690.

Harding, D., and A. Pagan. 2006. Synchronization of cycles. *Journal of Econometrics* 132: 59–79.

IMF (International Monetary Fund). 2002. *Recessions and recoveries. IMF world economic outlook, April*. Washington, DC: IMF.

Kedem, B. 1980. *Binary time series*. New York: M. Dekker.

Kedem, B. 1994. *Time series analysis by higher order crossings*. Piscataway: IEEE Press.

Mintz, I. 1969. *Dating postwar business cycles: methods and their application to Western Germany, 1950–1967. Occasional Paper No. 107*. New York: NBER.

Mintz, I. 1972. Dating American growth cycles. In *The business cycle today*, ed. V. Zarnowitz. New York: NBER.

Moore, G., and V. Zanrowitz. 1986. The development and role of the National Bureau of Economic Research's business cycle chronologies. In *The American business cycle: Continuity and change*, ed. R. Gordon. Chicago: University of Chicago Press.

National Centre for Econometric Research. MBBQ Code. Online. Available at http://www.ncer.edu.au/data/. Accessed 8 Sept 2006.

NBER (National Bureau of Economic Research). 2003. The NBER's recession dating procedure. Online. Available at http://www.nber.org/cycles/recessions.html. Accessed 25 Aug 2006.

OECD (Organisation for Economic Co-operation and Development). Leading indicators. Online. Available at http://www.oecd.org/std/cli. Accessed 25 Aug 2006.

Sichel, D. 1994. Inventories and the three phases of the business cycle. *Journal of Business and Economic Statistics* 12: 269–277.

Sichel, D. 1993. Business cycle asymmetry: A deeper look. *Economic Inquiry* 31: 224–236.

Stock, J., and M. Watson. 1991. A probability model of the coincident economic indicators. In *Leading economic indicators: New approaches and forecasting records*, ed. K. Lahiri and G. Moore. Cambridge: Cambridge University Press.

Watson, M. 1994. Business cycle durations and postwar stabilization of the U.S. economy. *American Economic Review* 84: 24–46.

Watson, M. Gauss code for Bry Boschan Algorithm. Online. Available at http://www.wws.princeton.edu/mwatson. Accessed 8 Sept 2006.

Business Cycles

Michael Dotsey and Robert G. King

Development of rational expectations models of the business cycle has been the central issue on the macroeconomic research agenda since the influential analyses of Robert Lucas (1972a, b). In this essay, we review these developments, focusing on the extent to which the rational expectations perspective has generated a new understanding of economic fluctuations.

Economists have long suspected that expectations play a central role in the business cycle, particularly in determining the relationship between money and economic activity. For example, Haberler's (1937) classic interwar survey of business cycle theory stresses the role of expectations, in a variety of theories that explain the business cycle as a Frischian (1937) interaction of external shocks and propagation mechanisms. Expectations also constitute an independent source of shocks in 'psychological' theories of the business cycle. However, as Haberler's survey makes clear, there has long been substantial disagreement among economists about the relative importance of various economic factors – sources of shocks and propagation mechanisms – in determining the observed character of business fluctuations. With the development of formal econometric analyses of business cycles – beginning with Tinbergen's work (Tinbergen 1932) and proceeding through Sargent (1981) – it has become clear that unrestricted models of expectations preclude a systematic inquiry into business fluctuations.

The postulate that expectations are rational in the sense of Muth (1961), i.e. that economic agents accumulate information and utilize information efficiently, imposes considerable discipline on business cycle analysis. At present, no single rational expectations model has captured all of the central elements of the business cycle. One could take the view that an ultimate explanation of economic fluctuations will require a return to 'psychological influences'. We prefer to believe that existing individual models highlight specific features that are important and that the gradual accumulation of knowledge about shocks and propagation mechanisms will ultimately yield rational expectations models consistent with observed business cycles.

The organization of our discussion is as follows. First, we briefly consider a set of 'stylized facts' that any successful model must minimally produce. Then, we turn to four categories of rational expectations models of the business cycle, considering in turn how each has been developed to account for some specific set of stylized facts. We then review the empirical evidence regarding the overall performance of each class of models.

We begin by exploring the role of expectations in the basic real business cycle models of Kydland and Prescott (1982) and Long and Plosser (1983), in which dynamics of business cycles reflect the interaction of temporary real shocks and intertemporal (capitalistic) production. We then consider the monetary business cycle models of Lucas (1972, 1973) and Barro (1976, 1980) which utilize incomplete information as a rationale for temporary real effects of monetary disturbances. Although agents have rational expectations in these models, lack of timely information on monetary shocks implies that agents erroneously perceive price level movements as representing changes in relative prices. After considering equilibrium models of the business cycle – in which prices are flexible – we turn to Keynesian models of business fluctuation constructed under the rational expectations postulate. Our discussion begins with the analyses of Fischer (1977) and Gray (1976), who model temporary wage stickiness arising from nominal wage contracts. Subsequently, we consider the emerging class of theories that focus on commodity price stickiness, beginning with a parable told by McCallum (1982) and then considering some alternative formal developments by Rotemberg (1982), Mankiw (1985), and Blanchard and Kiyotaki (1986).

Throughout our discussion, we follow the traditional macroeconomic practice of considering business cycles – defined as the stochastic components of macroeconomic time series – as

stationary stochastic processes. This practice is followed in our description of stylized facts, but is also implicit in the theoretical economies that we consider, since the time series generated by these economies are stationary. If, in fact, economic time series exhibit nonstationarity, as argued by Nelson and Plosser (1982), then these classes of models are called into question. In a concluding section we briefly discuss the ongoing development of rational expectations business cycles that are capable of producing model economies that have nonstationary components.

Stylized Facts

Much of our survey deals with the ability of various business cycle models to generate time series whose properties are consistent with commonly discussed summary statistics, i.e. the stylized facts of business cycles (see e.g. Lucas 1977). Presentations of these stylized facts typically proceed as follows. First, certain smooth curves are removed from the data, frequently after a logarithmic transformation; these eliminate deterministic growth and seasonal components. Summary statistics are then calculated on the transformed data.

At a minimum the list of real quantity variables to be considered consists of the major national accounts aggregates – consumption, investment and output – along with measures of labour input (manhours, employment). In addition, real wages, real money balances and certain financial activity variables are frequently considered, as in the growth rate of some nominal variables such as the money stock, nominal interest rates and prices. All of the quantity series – including real balances – exhibit significant positive serial correlation at the annual or quarterly interval. They all also display positive covariation, both with output and with each other. They differ somewhat in relative volatilities, notably investment is more volatile than output, which in turn is more volatile than consumption. Evidence concerning the cyclical behaviour of the real wage is inconclusive; in part, this reflects a variety of constructs used. In general, however, there does not appear to be a pronounced cyclical relation. Measures of financial activity – such as deposit turnover and bank clearings – are strongly procyclical (Mitchell 1957). As Lucas (1977) observes, there is little reason to qualify the observations by reference to specific time periods.

However, the relationship between nominal variables and the cycle exhibits less stability over time. In Mitchell's (1951) consideration of interwar data for the US, the price level and short-term nominal interest rate were strongly procyclical. More recent investigations by Hodrick and Prescott (1980) into post-war US cycles, document a changing relation, price levels are countercyclical during the latter half of their sample and short-term rates are not systematically related to economic activity. However, most investigations do document a positive relation between income velocity and real activity that mirrors the financial transactions data.

When many sectors are included in this analysis, as in Mitchell (1951), there is a tendency for co-movement across sectors and considerable stability in lead–lag relations relative to aggregate output. There do appear to be different degrees of sectoral co-movement and amplitude. For example, agriculture does not covary closely with the rest of the economy. Producer and consumer durable goods manufacturing exhibits greater volatility than services.

Expectations and Real Business Cycles

In recent years, macroeconomists have begun the long postponed task of developing basic equilibrium models of economic fluctuations. That this is an essential first step was cogently argued by Hicks (1933) over fifty years ago, who stressed that one could not measure the extent of disequilibrium without first determining the content of equilibrium theory and that, in a dynamic stochastic system, there is rich content to equilibrium theory.

The analyses of Kydland and Prescott (1982) and Long and Plosser (1983) explain the dynamics of business cycles as reflecting the interaction of real shocks – to total factor productivity – and intertemporal (capitalistic) production possibilities. The Long and Plosser (1983) analysis develops

some general economic principles – mentioned by Haberler (1937) – by studying the decisions of a representative consumer ('Robinson Crusoe') who directly operates the production technology of the economy. In this context the business cycles that arise are Pareto efficient. Thus, the mechanisms that generate cyclical activity are quite general and should carry over to richer macroeconomic models that possess incomplete information and nominal rigidities, including those that we consider below.

For example, the analysis of Long and Plosser shows that even if disturbances to production possibilities are temporally independent, real quantities – output, consumption and capital – display positive serial correlation. Shocks are propagated over time due to the preference of economic agents for smoothing consumption, and the fact that the intertemporal technology makes smoothing feasible. However, the persistence of shocks is limited by the existence of fixed factors (such as a fixed endowment of time); the ultimate effects of shocks are negligible. That is, in the periods after a productivity shock, there is negative net investment – relative to a trend value – as the economy adjusts back towards a steady-state. This residual role of investment implies that it displays great cyclical volatility (see King and Plosser 1986). Thus, with temporally independent shocks, the basic equilibrium model predicts some of the central stylized facts – positive serial correlation in consumption and production, as well as the relative volatilities – but fails to capture the positive serial correlation of investment.

When there are many commodities, Crusoe's preference for diversity in his consumption bundle means that the effects of a temporary productivity shock in one sector are also transmitted across other sectors. Thus, as Long and Plosser (1983) stress, the basic equilibrium model also predicts that there will be comovement, with economic activity in diverse sectors tending to rise and fall together. Therefore, the basic equilibrium model also predicts another of the centralized stylized facts emphasized by Mitchell (1951).

Temporary shocks to factor productivity typically exert offsetting income and substitution effects on the effort decision, so that Crusoe's optimal cyclical variation in employment is ambiguous. In the parametric models of Long and Plosser, these two effects offset exactly, so that there is zero cyclical variation in labour input. Kydland and Prescott (1982) explore the implications of greater intertemporal substitution in preferences, using a non-time-separable but recursive preference specification. In this case, Crusoe finds it optimal to substitute effort toward periods in which its marginal reward is high, which leads effort to respond positively to temporary productivity shocks.

Even with temporary shocks to productivity, the intertemporal consequences of capital accumulation and effort decisions imply that Crusoe must form expectations about future production opportunities. In general, optimal decision rules will be different when Crusoe makes alternative assumptions about the nature of shocks and for alternative specifications of preferences and technology. Furthermore, the rational expectations assumption plays a pivotal role in the process of transforming optimal social decisions into a competitive theory of fluctuations, for there will be a coincidence between Crusoe's (or a social planner's) decisions and the decentralized actions of private agents only if expectations are rational.

Expectations have additional implications for Crusoe's decision rules when there is serial correlation in the exogenous factors, i.e. total factor productivity. For example, Crusoe's incentives for saving/investment to achieve consumption smoothing, are reduced if changes in (expected) future productivity accompany changes in current productivity because of larger wealth effects of such changes. Further, anticipated future variations in productivity also affect the marginal reward to current investment and the rewards to future effort, exerting additional substitution effects on current decisions.

Evidence on Real Business Cycles

Although real business cycle models produce some qualitative features of the business cycle it remains to determine whether they explain

fluctuations *quantitatively*. The initial research effort addressing these questions has been undertaken by Kydland and Prescott in an influential series of papers (summarized in Prescott 1986).

Following the methodological recommendations of Lucas (1980), Kydland and Prescott restrict the number of free parameters in their model economy by a number of steady-state conditions and also by the extensive use of behavioural parameter estimates taken from applied studies in other fields. For example, they use the observed constancy of labour's share to pin down the parameters in a Cobb–Douglas production function, and results from analyses of financial markets to restrict a preference curvature parameter governing the extent of intertemporal substitution/risk aversion. Following Solow (1957), they measure variations in total factor productivity as a residual from the aggregate production function and choose a simple Markovian stochastic process to capture the serial correlation in this series.

The results of the Kydland–Prescott studies have been surprising to most economists. The initial model economy produced summary statistics – second moments of consumption, investment, output, productivity, and effort – that accorded with the stylized facts described previously. (The specific presentation of the stylized facts to which the Kydland–Prescott model was compared is contained in Hodrick and Prescott 1980). However, it is also clear that the basic neoclassical business cycle model as developed by Kydland and Prescott does not meet the stringent standards of rational expectations econometrics. Altug (1985) subjects the Kydland–Prescott model to rational expectations econometric procedures and finds that the model's restrictions are rejected by the data. Given the level of abstraction currently found in this model this is perhaps not surprising; it is encouraging that these types of models can loosely mimic some important aspects of cyclical activity.

The basic neoclassical model of Kydland and Prescott has been criticized on a number of other grounds that warrant further discussion. First, the model has no implications for any cyclical variation in employment or unemployment. That is, the model uses the representative agent paradigm and permits a smooth tradeoff between hours and output, so all adjustments in labour effort take place in terms of hours and not numbers of workers. Forcing a (more realistic) choice between working full time or not at all would generally introduce problematic nonconvexities in production possibilities. However, the important work of Rogerson (1985) provides a method for analyzing production nonconvexities in a representative agent model. Rogerson uses the fact that by introducing social arrangements that formally resemble lotteries – in that they specify probabilities of working full time or not at all – the representative agent problem can be made convex. This contrived randomness in the representative agent and corresponding social planner's problem improves welfare by smoothing the opportunities for effort by averaging across the population. It corresponds in a competitive, multi-agent framework to an economy in which some agents are employed and some are not, and in which their relative numbers can fluctuate over time. Further, the indivisibility of work effort results in a dramatic change in the corresponding social planner's problem that can be used to compute competitive outcomes. This optimum problem can be interpreted as that of a single agent with a greater degree of intertemporal substitution in labour supply than that of the identical agents that populate the economy. This is empirically important within the Kydland–Prescott model, since the greater degree of intertemporal substitution in aggregate labour supply is capable of producing quantitatively greater variations in employment (see Hansen 1985).

The main potential theoretical alternative for smoothing production nonconvexities is to allow for heterogeneity of preferences. However, given the primitive state of methods for solution and estimation of dynamic macroeconomic models this alternative is not practical at the moment. It is, therefore, likely that Rogerson's insight will be widely employed. Overall, the focus of most business cycle models on hours and not on the number of employed workers represents a transient feature and is not an essential character of the real business cycle approach.

The social criticism of the Prescott model is that its internal mechanisms do not by themselves produce much serial correlation in economic time series. This is because, even though the model provides a mechanism for the propagation of shocks, the share of physical capital in output is small (about one-third). Therefore, consumption smoothing cannot be very important quantitatively in this framework (see King and Plosser 1986). Rather, the cyclical character of the variation in total factor productivity – the Solow residual which is a Markov process that is close to a random walk – is used to generate persistence.

The stochastic nature of the shocks is therefore a key ingredient for generating the cyclical behaviour in the Kydland–Prescott model and there has also been some scepticism directed toward the nature of these shocks. For example, one questions whether this construct really captures an exogenous variable (technological change). If cyclical variations in the intensity of utilization of capital and labour inputs are significant, then important biases could arise, since endogenous decisions with respect to utilization will incorrectly be attributed to changes in technology. Further, in industries that are noncompetitive, there may be cyclical variations in the relationship between marginal cost and price (mark-ups) that would be counted as shocks to factor productivity by the Solow–Prescott procedure (see Bils 1985). Also, Barro (1986) and others have expressed scepticism that there are real shocks of sufficient magnitude to generate observed cycles.

Finally, with the exception of King and Plosser (1984), these models cannot generate any of the observed correlations between money and economic activity, since financial sectors have been omitted from most real business cycle models.

King and Plosser (1984) extend the real business cycle model by incorporating accounting services as a factor in production of final goods. Consequently, when there are increases in total factor productivity in the final goods sector, there is an induced increase in the quantity of such services (an intermediate good), which rationalizes Mitchell's (1951) finding that measures of transactions activity in the banking sector are strongly procyclical. In considering extensions to incorporate demand deposits and outside currency, King and Plosser follow standard macroeconomic practice by assuming that service flows are proportional to asset stocks. Therefore, real quantities of currency and demand deposits covary positively with economic activity. Moreover, if price levels are not too countercyclical, then nominal demand deposits will move with the cycle, while movements in nominal currency may be unrelated to the evolution of the cycle, a hypothesis for which King and Plosser provide some supporting empirical evidence. However, as McCallum (1986) points out, if the central bank is targeting currency plus deposits, then these correlations can also arise in a monetary business cycle.

The main contribution of this literature is the detailed development of propagation mechanisms which may not be sensitive to the nature of the shock. Therefore, the real business cycle literature may serve as a useful complement to other equilibrium business cycle models, such as those involving monetary impulses. It is to this class of models that we now turn.

Money, Expectations and Business Cycles

The pioneering work incorporating rational expectations into monetary models of the business cycle was undertaken by Lucas (1972a, b, 1973). Macroeconomist's concern with linking the real and monetary sides of the economy probably stems from the influential work of Friedmand and Schwartz (1963), which appears to document an important causal role for nominal impulses including shifts in the money supply and the velocity of circulation.

The basic feature of imperfect information variants of equilibrium business cycle theory can be depicted in a simple log-linear business cycle model that essentially follows Lucas (1973). In this model, a non-storable commodity is produced at distinct locations indexed by z. Production in each location depends linearly on last period's output and on the perceived relative price, $p_t(z) - E_z p_t$, where $E_z p_t$ is the expected value of the log of the aggregate price level.

Output demand at any location is positively related to factors influencing aggregate demand and a relative demand shock.

To close the model, one must specify a stochastic process governing the supply of money and the information set available to agents at each location. Agents are typically assumed to know the economy's structure, their current local price, $p_t(z)$, and past values of all variables and disturbances. They do not observe the contemporaneous values of aggregate data or of the disturbances.

This simple framework yields a number of key results that extend to other members of this class of equilibrium business cycle models. The primary result is that it is only *unperceived* monetary disturbances which produce real effects. Perceived changes in money affect both local and aggregate prices uniformly so that these are neutral toward relative prices and real activity. It is instructive to trace through the effects of a positive monetary shock. The demand for goods at location z rises, causing an increase in the price at location z. With incomplete information, suppliers in location z do not know whether any particular increase in $p_t(z)$ such as that arising from the monetary shock is due to aggregate or relative disturbances. Given the stochastic structure of the model, agents will generally attribute some of a money induced movement in $p_t(z)$ to an improvement in relative prices and therefore they will supply more. (The proportion of the price movement attributed to relative shifts in demand depends on the underlying variances of two shocks.) Therefore, an *unanticipated* increase in the money supply will cause output to rise precisely because it is mistakenly perceived as representing a change in relative prices. If, on the other hand, agents accurately perceived the shift in the money supply, they would neutralize the effects of this disturbance. Sargent and Wallace (1975) use this to develop the implication that anticipated movements in money supply have no real effects.

An initial criticism of Lucas's analysis involved the fact that this simple model could not generate the serial correlation evident in economic time series (Hall 1975). But, as Lucas (1975) argues, linking the model of monetary shocks to capital accumulation and the other propagation mechanisms of real business cycle theory potentially overcomes this difficulty. For example, Sargent (1979) provides a nicely worked out linear business cycle model that utilizes adjustment costs to propagate temporarily misperceived nominal shocks.

The neutrality of perceived monetary disturbances represents a substantial problem for this class of equilibrium business cycle models. In reality, monetary data (although somewhat noisy) is produced in a very timely manner. If the relevant decision period is approximately one quarter, agents' information sets should plausibly be modelled as including the available contemporaneous monetary data. In this situation, King (1981) shows that fluctuations in output should be uncorrelated with the reported monetary statistics, essentially because expectation errors about relative prices should be uncorrelated with available information. Further, revisions in the monetary statistics should be correlated with real activity because the initial reporting errors induce misperceptions.

Thus, if monetary disturbances are accurately perceived, then they cannot be business cycle impulses for the reason suggested by Lucas (1972, 1973). It is important to stress that this does not rule out incomplete information as a rational for the non-neutrality of other nominal disturbances (such as money demand shocks) that may more plausibly be not directly observable over the relevant decision period.

Moreover, King's (1981) result is conditioned by the assumption that monetary disturbances are exogenous. If the central bank leans against changes in interest rates or if changes in inside money are correlated with real activity, then contemporaneous monetary statistics may be correlated with output even if they are accurately perceived. Moreover, King and Trehan (1984) show that monetary shocks can be non-neutral due to a signalling effect, if these statistics convey information about unobservable real economic conditions that influence agent's production and investment decisions.

It has also been suggested that King's result may be too strong, since although monetary data

is available it may also be quite costly to process. Therefore, agents may in some sense ignore the data in making their labour/leisure decisions, which would imply that the initial specification of the information set was appropriate. (Edwards (1980) constructs a model in which there is a competitively determined fraction of agent that acquire costly information about the true monetary state, but it is unclear from his analysis whether business cycles can be a large social problem if the individual costs of information are small.) The preceeding argument reveals the arbitrary manner in which information structures are specified in this class of models and this is a problem that has not been dealt with satisfactorily in the macroeconomics literature to date.

There are numerous extensions and modifications of the simple model just considered. The most notable are those of Barro (1976, 1980), which are motivated by intertemporal substitution possibilities rather than by contemporaneous expected relative prices (as in Lucas 1973; Friedman 1968). But these analyses preserve the central empirical implications of the simple model: (i) the irrelevance of predictable variations in monetary policy; and (ii) the causal link between unperceived monetary disturbances and real activity.

Empirical Analyses of Money and Business Cycles

The empirical work on monetary impulses in equilibrium business cycle models is much too extensive to cover completely in this essay. Rather, we review three major lines of empirical investigation that bear on the relevance of the line of research. By and large, the evidence suggests that models of this class do not adequately represent links between money and business cycles.

Tests based on monetary decompositions. The first layer of tests examined the relationship between unanticipated movements in nominal variables and economic activity, with the key references being Sargent (1973, 1975) and Barro (1977, 1978). Following Barro's lead, subsequent investigations have focused on reduced form relations between money and economic activity, rather than estimation of systems incorporating a 'Lucas supply function' as in Sargent's early studies. The idea behind the Barro-type tests is to decompose the observed monetary time series into unanticipated and anticipated components by specifying a prediction rule. This two stage procedure involves estimation of a money supply process, with the residuals treated as unanticipated money and the fitted values treated as anticipated money. The empirical studies then investigate whether constructed unanticipated money influences various measures of economic activity and if the constructed anticipated components of money are neutral. Initial tests by Barro utilized a two-step procedure, with later investigations employing the econometrically more efficient method of estimating a simultaneous equation system and testing cross equation restrictions (Leiderman 1980; Abel and Mishkin 1983).

These tests concern the joint hypothesis that expectations are rational, that the money supply process is correctly specified, that the process governing the behaviour of the economy is correct, and that anticipated money is neutral. Thus, correct specification of all of these elements is necessary for successful execution of these tests. For example, if the Federal Reserve's reaction function is misspecified through the exclusion of relevant variables then measures of anticipated money will include the effects of these variables. If these excluded variables are correlated with explanatory variables in equations that depict the behaviour of the relevant economic magnitudes under consideration, which is likely to be the case, then coefficients will be biased and test statistics will be inappropriate.

The results of this type of tests are mixed. The analysis of Barro (1977) concerning the relationship between money and unemployment supports the implications of equilibrium business cycle theory. Working at the annual interval, Barro provides evidence that (i) anticipated monetary changes do not affect real activity in a statistically significant manner; and (ii) that unanticipated money growth affects output over three years, with the peak effect concentrated in the second year. A follow-up study of the price level at the

annual interval, Barro (1978), provides evidence that price level movements accord less well with the predictions of theory. Although anticipated monetary changes have a one-for-one impact on the price level, the response of the price level to monetary shocks is more protracted than the response of real activity. Barro and Rush (1980) provide additional evidence using data on unemployment, output, and prices from the quarterly post-war time series, the interval that has subsequently been studied by most researchers. Generally this study confirms Barro's earlier results that unanticipated money influences real GNP (positively) and unemployment (negatively) but, as with the annual data, the results involving the price level are less persuasive. Although unanticipated money does affect the price level less than one for one, the lag structure for unanticipated money is inconsistent with lags found in output and unemployment equations.

Working at the quarterly interval, Mishkin (1982) and Merrick (1983) provide evidence against the neutrality hypothesis, where the hypothesized money supply process and lag lengths are altered from the Barro–Rush specification. Merrick essentially tries to replicate the Barro–Rush quarterly results on real GNP, after altering the money supply process by including lagged Treasury bill rates and stock market returns. He finds that unanticipated money no longer affects real GNP, but that anticipated money does. Mishkin also alters the money supply process by including past Treasury bill rates but finds that this does not affect the Barro-Rush results over a somewhat different sample period, where an eight quarter maximum lag is imposed. However, upon extending the lag lengths on unanticipated and anticipated money to twenty quarters, he is able to reject the joint hypothesis of rationality and neutrality. The Merrick and Mishkin results cast doubt on the robustness of the neutrality results obtained at the annual interval. However, in interpreting the above results, one must keep in mind that a composite hypothesis is being tested. For example, if anticipated money was neutral, but if the central bank engaged in interest-rate smoothing – as in Goodfriend (1986) – then variations in money growth would accompany changes in the real interest rate. If the factors that lead to these changes in the real interest rate are omitted in the output equation, anticipated money will spuriously appear to be non-neutral.

Leiderman (1983) investigates the cyclical pattern of real wage movements in response to money on both annual and quarterly data. According to neoclassical theory, the real wage should decline with application of an increased amount of effort to a fixed stock of capital. Thus, if misperceived monetary shocks fool labour suppliers into working more, then monetary shocks should lower real wages and increase output, so that a countercyclical relationship emerges between monetary shocks and real wages. Also, predictable shifts in money will leave real wages unaffected. Leiderman finds some support – at both the annual and quarterly intervals – for countercyclical variation in the real wage, which is strongest when the real wage is deflated by the wholesale price index and when overtime payments are excluded. However, in a recent study of a number of manufacturing industries, Kretzmer (1985) finds evidence that industry specific product wages (industry wage divided by the industry wpi component) are uniformly positively related to unanticipated monetary shocks.

Granger causality tests. Another type of neutrality test is based on the following observation: given the relevant state of the economy (capital, etc.), the history of monetary shocks should have no effects on real activity. Sargent (1976) and Sims (1980) utilized this perspective to construct neutrality tests along Granger causality lines. In a multivariate context nominal variables should not Granger-cause (predict) a vector of real variables if these contain the economy's state variables. (Conditions that assure that the state variable is reputable in this form are provided by Sargent (1979) – some may be unwilling to impose such lag length restrictions on error terms, which Sims (1980) argues incredible.) Sargent (1976), Sims (1980) and Eichenbaum and Singleton (1986) illustrate that the results of such tests are heavily dependent on variable selection and data processing, particularly treatment of non-stationarities.

A variant of this procedure is employed by Haraf (1983), who examines a four variable vector autoregression using real output, employment, inventories, and backorders. A constructed unanticipated money series does not Granger-cause the vector process governing the four real variables in the model, a result that is consistent with the simple equilibrium business cycle model. However, Haraf also finds that with the exception of real GNP, contemporaneous unanticipated movements in money have little explanatory power once lagged model variables are taken into account.

Tests based on contemporaneous monetary data. The previous tests concentrated on the distinction between unanticipated and anticipated changes in money. However, equilibrium business cycle theory typically predicts that the relevant distinction is between perceived and unperceived movements in money. Since monetary statistics are readily available, agents misperceive the true monetary state of the economy only to the extent that monetary statistics contain some reporting errors. Therefore, revisions in monetary statistics are indicators of misperceived money, and it is misperceived money that should be the relevant variable in explaining real economic fluctuations. Specific tests of the equilibrium business cycle theory using contemporaneous monetary data – historical statistical reports that were *potentially* available to private agents – are conducted by Barro and Hercowitz (1980) and Boschen and Grossman (1982).

Both of these papers contain evidence contradicting the implications of the simple equilibrium business cycle model outlined above. Barro and Hercowitz find that revisions in the monetary data do not help explain cyclical fluctuations of output or unemployment. Boschen and Grossman focus on King's (1981) observation that output should be uncorrelated with available monetary data. They begin by constructing a more elaborate procedure that yields valid tests of the real effects of exogenous perceived money on output when misperceived money can affect output through a specific propogation mechanism. They find that contemporaneous monetary data is significantly (partially) correlated with real activity, which is inconsistent with the theory. Boschen and Grossman also test whether monetary reporting errors have real consequences and as in Barro–Hercowitz, there is no evidence of real effects. Thus, the Boschen and Grossman findings are inconsistent with the joint hypothesis of (i) a specific equilibrium business cycle model; (ii) that agents utilize contemporaneous information as money; and (iii) that measures of money (original and final reports) are exogenous.

Although properly specified tests are difficult to conduct, the mixed results of these three types of tests does not provide strong support for the equilibrium monetary business cycle view. Consequently, investigation of Keynesian alternatives seems warranted. We begin with the notion that multiperiod contracting imparts some stickiness to the nominal wage.

Nominal Wage Contracting Models

Much of the nominal wage contracting literature is based on two lines of work. One originates in Taylor (1979, 1980) and the other follows from Gray (1976) and Fischer (1977).

Taylor (1979, 1980) develops a model with multiperiod, overlapping nominal wage contracts and mark-up pricing. Simulations of the model under the assumption that wage contracts last for three or four quarters are used to investigate the dynamics of output or unemployment. Without any of the neoclassical propagation mechanisms, Taylor's models generate substantial serial correlation from the interactions of wage setting rules and expectations – shocks can last for more than the contract length because these are passed along via other, subsequent contracts. But Taylor's models have been criticized as departing too far from wage setting rules that could plausibly be rationalized by neoclassical methods – thus involving wage setting based on predetermined wage rates of others, which should be irrelevant – and for not containing the natural rate property (for further discussion of Taylor's models, see McCallum 1982).

The Gray (1976) and Fischer (1977) perspective on wage contracts can be developed as

follows. Production takes place at various locations or industries indexed by z, and depends negatively on the real wage $w_t(z) - p_t(z)$ in each location. (All variables are expressed in logarithms.) In the one period ahead contracting version of the model, the nominal wage $w_t(z)$ is set according to the rule $w_t(z) = E_{t-1}p_t + \gamma(z)(P_t - E_{t-1}P_t), \gamma(z)$ indicates the extent of indexing in industry z. If $\gamma(z) = 1$, then wages in z are completely indexed to the aggregate price level. Given the nominal wage, firms determine employment along their marginal product curve, the efficiency condition being that the marginal product of labour equals $w_t(z) - p_t(z)$. Therefore a rise in the real wage reduces employment and output at location z.

Aggregate demand at any location is directly related to aggregate real balances and a relative demand shock, as in the equilibrium business cycle model. Also, the money supply is assumed to follow a random walk. In this setting, with incomplete indexing ($\gamma(z) < 1$), a positive money supply shock causes real wages to fall and output to rise. Also, with contracts set at one period in length, shifts in money that were anticipated at $t-1$ have no real effects. Therefore, tests that only consider the distinction between anticipated and unanticipated money cannot distinguish between equilibrium business cycle models with no contemporaneous information and models with nominal contracts extending for only one period.

However, as Fischer (1977) indicates, when contracts last for more than one period, shifts in money that are anticipated at $t-1$ will have real effects since some locations are locked into contracts conditioned on period $t-2$ information. However, Fischer (1980) reports some difficulties in implementing this strategy.

A direct test of the contracting model is performed by Ahmed (1986. Ahmed undertakes a careful study of the relationship between the Phillips curve slope and the degree of wage indexation in a particular industry. (The data set includes 19 Canadian industries.) The contracting model predicts that the responsiveness of industry specific output to unanticipated changes in money should be inversely related to the degree of indexing. That is, greater indexation by a particular industry reduces the responsiveness of real wages to unanticipated money and reduces the change in industry output to a monetary disturbance. Ahmed finds no evidence that there is any relationship between indexation and the magnitude of responsiveness of industry specific output to an aggregate monetary shock. These results are at variance with the implications of the contracting model.

Therefore, the strategy of producing monetary business cycles through nominal wage rigidities does not receive strong empirical support. This has lead Keynesians to refocus their attention on nominal rigidities that may occur in other areas of the economy, namely in the price of specific commodities.

Sticky Prices and Business Cycles

After the Dunlop–Keynes–Tarshis controversy of the 1930s unveiled the lack of confirmation for countercyclical real wages, Keynesian macro theorists turned from models incorporating stickiness of wages to models featuring stickiness of product prices. This activity spanned the range from rationalizations of the pricing equations in large scale econometric models to the abstract dynamic pricing model of Phelps and Winter (1970) and the non-market clearing theory of Barro and Grossman (1976). Curiously, this prior path seems to have been ignored by the profession at large. Until recently, there has been substantial effort allocated to sticky wage models despite their reliance on a countercyclical path for the real wage. However, the past several years have seen increased attention to sticky price models. Although this line of research is still at an early stage and has, as yet, generated little empirical literature, we provide a brief review because of its likely importance in coming years.

Simultaneously with Fischer's wage contract model, Phelps and Taylor (1977) propounded a basic rational expectations model with price stickiness, in a paper that has received far less professional attention than Fischer (1977). However, research into sticky price models was continued

by McCallum in an important series of papers. Initially, McCallum focused his investigations on the conditions under which sticky price models rationalized non-neutrality of monetary shocks while maintaining the neutrality of anticipated monetary policy (1978, 1979, 1980).

More recently, McCallum (1982, 1986) has provided a detailed outline of interactions between nominal shocks, price adjustment, and real activity, which presumably will be developed further in coming years. The key elements of this story are as follows. To economize on certain costs, firms find it optimal to maintain a set nominal price over some period, accommodating variations in relative and aggregate demand through alterations in production and inventories. Thus, monetary shocks have real effects. However, price adjustments incorporate firms' anticipations about monetary policy, so the real consequences of anticipated movements in money are much smaller than unanticipated movements and may be fully neutralized.

In McCallum's work the period over which stickiness prevails plays a crucial role. If price stickiness is to be assigned a major role in business cycles – even as an impulse mechanism – then the period over which firms elect to make prices sticky must be non-trivial. McCallum (1982, 1986) begins by reviewing theoretical explanations of why producers might temporarily stabilize relative prices against shocks, for example to attract a clientele of customers who prefer relative price stability. He then argues that the costs within period adjustment of nominal prices – or of indexation that would neutralize monetary shocks – cannot be the physical costs of adjusting prices, but rather are computational costs associated with the difficulties that agents face in understanding more complex contracts. He also argues that indexation provides only small reductions in risks to participants, although it is unclear how this is consistent with business cycles that are an important social problem.

Some other recent attempts to give theoretical content to the idea of price stickiness have proceeded along two different paths. One avenue emphasized by Mankiw (1985) and Blanchard and Kiyotaki (1985) involves models with monopolistically competitive firms that face fixed ('menu') costs of adjusting prices. So far, this line of research has concentrated on establishing that menu costs that are small can lead to large departures from socially efficient allocations when nominal shocks occur. These models are not yet dynamic, so that distinctions between anticipated and unanticipated movements in nominal variables have not yet been explored. But it stands to reason that there would be results that differed from McCallum's, since in his setup there are effectively zero costs of adjusting prices between periods and infinite costs of changing prices within the period. First, as in Mankiw (1985), large nominal shocks – even if unanticipated – would tend to be neutralized. Second, small anticipated changes in money would tend not to be neutralized, as the menu costs would be prohibitive. Irrespective of one's view on the plausibility of menu costs, these recent analyses provide a clue as to how individual agents might regard the gains to altering nominal contracts as small even though the social benefits would be large, due to the suboptimality of monopolistically competitive equilibria.

Another line of research has been pursued by Rotemberg (1981), who employs quadratic costs of price adjustment to induce gradual price adjustment. As in Phelps–Winter, these costs are viewed as arising from an erosion of the firm's clientele, with a specific interpretation involving an individual's dislike of price volatility. Using rational expectations methodology, Rotemberg provides evidence that prices adjust gradually, although the specific structural models which he employs are inconsistent with the cross equation constraints implied by the rational expectations postulate.

As the dynamic implications of sticky-price macro models are developed in more detail, it will become possible to discriminate between these models and the flexible price equilibrium theories considered earlier. In this process, since price level behaviour is a result of the interaction between private agents and the monetary authority, an adequate definition of price stickiness will

be required. In particular most researchers have focused on the smoothing of price level variations that arises from private sector actions. However, smoothing can also arise from systematic actions by the monetary authority (see Goodfriend 1986). Powerful tests will presumably require systematic examination of data generated prior to the creation of the Federal Reserve.

The microeconomic evidence developed by Carlton (1986) – working with the Stigler and Kindahl (1970) data – shows that some prices are fairly rigid. However, the rigidities do not seem to conform to those that have been postulated by macro-modellers. For instance, many price changes are extremely small, indicating that menu costs are not a pervasive factor. Carlton also does not find much evidence that buyers have strong preferences for products whose prices are relatively stable, implying that one rationalization of Rotemberg's costs of adjustment is apparently inoperative. As the particular mechanism that generates rigidities could be quite important for the dynamic implications of this class of models, identification of the empirically relevant sources of rigidities is necessary. At this stage, this class of models should be regarded as a potentially promising means of resurrecting long standing Keynesian notions. As of yet their value has not been proven.

Conclusion

In our overview of rational expectations models of business fluctuations, we have consciously emphasized the extent to which this class of models has generated cyclical interactions that are consistent with empirical evidence. Evidently, progress has not been rapid and there is currently no compelling evidence for any particular description of cycles, despite the fact that the models quite frequently have substantially distinct policy implications. We do not regard this assessment as a reason for departing from the discipline imposed by rational expectations, but feel that this is rather an indication of the amount of work that remains to be done.

In fact, some recent research has led us to become less sure that the conventional representation of business cycles – the stochastic components of economic time series – is appropriate. Nelson and Plosser (1982) have produced some provocative empirical work which cannot reject the hypothesis that the stochastic components of economic time series are nonstationary, possessing random walk components. Although their tests have low power against the alternative that the stochastic components are stationary but highly persistent (McCallum 1986), these results represent a serious challenge to existing views. Further, there are now basic equilibrium models of fluctuations that imply non-stationarity if the intertemporal technologies are restricted so that the mean rate of economic growth is endogeneously determined (King and Rebelo 1986), basically because fixed factors are not too important. Further, these endogenous growth models have substantial implications for model building under the rational expectations postulate, for they imply that there are transformations of non-stationary economic variables that are stationary – that is, the macroeconomic data possess a cointegrated representation (King et al. 1986).

Our forecast is that the construction of rational expectations model of the business cycle will be the centrepiece of the macroeconomic research agenda over the next 15 years, as much as it has been over the 15 that have passed since Lucas's influential contributions (1972a, b). Recently, Lucas (1985a) has argued that economic fluctuations pale in welfare significance relative to the factors that determine the growth path of a particular country's economy; his research has recently turned to analyses of these factors (1985b). Most macroeconomists presumably share McCallum's (1986) scepticism that economic fluctuations are second order problems relative to economic growth and, hence, would doubt that Lucas's current research direction will have the impact of his 1972 work. But we are not so sure, for if the analysis of King and Rebelo (1986) is sustained in richer models, then it is inappropriate to separate the study of economic fluctuations from that of economic growth. That is, the fact that economies grow tells us that temporary shocks to the economy's production possibilities will have permanent effects on the level of output.

See Also

- ▶ Credit Cycle
- ▶ Depressions
- ▶ Multivariate Time Series Models
- ▶ Rational Expectations
- ▶ Trade Cycle

References

Abel, A.B.., and F.S. Mishkin. 1983. An integrated view of tests of rationality, market efficiency and the short-run neutrality of money. *Journal of Monetary Economics* 11: 3–24.

Ahmed, S. 1986. Wage indexation and the Phillip's curve slope: A cross-industry analysis. *Journal of Monetary Economics* 16: 69–88.

Altug, S. 1985. *Gestation lags and the business cycle: An empirical investigation*. Unpublished working paper, University of Minnesota.

Barro, R.J. 1976. Rational expectations and the role of monetary policy. *Journal of Monetary Economics* 2: 1–32.

Barro, R.J. 1977. Unanticipated money growth and unemployment in the United States. *American Economic Review* 67(March): 101–115.

Barro, R.J. 1978. Unanticipated money, output and the price level in the United States. *Journal of Political Economy* 86: 549–580.

Barro, R.J. 1980. A capital market in an equilibrium business cycle model. *Econometrica* 48: 1393–1417.

Barro, R.J. 1986. *Comments of Eichenbaum and Singleton*. Unpublished working paper, University of Rochester.

Barro, R.J., and H.I. Grossman. 1976. *Money, employment and inflation*. Cambridge: Cambridge University Press.

Barro, R.J., and Z. Hercovitz. 1980. Money stock revisions and unanticipated money growth. *Journal of Monetary Economics* 6: 257–267.

Barro, R.J., and M. Rush. 1980. Unanticipated money and economic activity. In *Rational expectations and economic policy*, ed. S. Fischer. Chicago: University of Chicago Press for National Bureau of Economic Research.

Bils, M.J. 1985. Real wages over the business cycle: Evidence from panel data. *Journal of Political Economy* 93: 666–689.

Bils, M. 1986. *Cyclical behavior of marginal cost and price*. Unpublished working paper, University of Rochester.

Blanchard, O.J. and N. Kiyotaki. 1985. *Monopolistic competition, aggregate demand externalities and real effects of nominal money*. National Bureau of Economic Research working paper No. 1770.

Boschen, J., and H.I. Grossman. 1982. Tests of equilibrium macroeconomics using contemporaneous monetary data. *Journal of Monetary Economics* 10: 309–334.

Carlton, D. 1986. *The rigidity of prices*. National Bureau of Economic Research working paper No. 1813.

Edwards, M. 1980. *Informational equilibrium in a monetary theory of the business cycle*. Unpublished working paper, University of Rochester.

Eichenbaum, M. and K. Singleton. 1986. *Do equilibrium real business cycle theories explain post-war U.S. fluctuations?* Unpublished working paper, Carnegie-Mellon University.

Fischer, S. 1977. Long term contracts, rational expectations and the optimal money supply rule. *Journal of Political Economy* 85: 191–205.

Fischer, S. 1980. On activist monetary policy with rational expectations. In *Rational expectations and economic policy*, ed. S. Fischer. Chicago: University of Chicago Press for National Bureau of Economic Research.

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 for National Bureau of Economic Research.

Frisch, R. 1933. Propagation problems and impulse problems in dynamic economics. In *Essays in honour of Gustav Cassel.* London: George Allen & Unwin.

Goodfriend, M.S. 1986. *Interest rate smoothing and price level trend-stationarity*. Unpublished working paper, Federal Reserve Bank of Richmond.

Gray, J. 1976. Wage indexation: A macroeconomic approach. *Journal of Monetary Economics* 2: 221–235.

Haberler, G. 1937. *Prosperity and depression*. Geneva: League of Nations.

Hall, R.E. 1975. Rigidity of wages and persistence of unemployment. *Brookings Papers on Economic Activity* 2: 301–349.

Hall, R.J. 1980. Comments on chapters 6 and 7. In *Rational expectations and economic policy*, ed. S. Fischer. Chicago: University of Chicago Press for National Bureau of Economic Research.

Hansen, G. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16: 309–328.

Haraf, W. 1983. Tests of a rational expectations–structural neutrality model with persistent effects of monetary disturbances. *Journal of Monetary Economics* 11: 103–116.

Hicks, J.R. 1933. Equilibrium and the cycle. Reprinted in J.R. Hicks, *Money, interest and wages: Collected essays in economic theory*, vol. 2. Cambridge, MA: Harvard University Press, 1982.

Hodrick, R.J. and Prescott, E.C. 1980. *Post-war U.S. business cycles: An empirical investigation*. Unpublished working paper, Carnegie-Mellon University.

Kertzmer, P.E. 1985. *Cross-industry tests of an equilibrium business cycle model with rational expectations*. Unpublished working paper, Board of Governors of the Federal Reserve System.

King, R.G. 1981. Monetary information and monetary neutrality. *Journal of Monetary Economics* 7: 195–206.

King, R.G. and C.I. Plosser. 1986. *Production, growth and business cycles*. Unpublished working paper, University of Rochester.

King, R.G., and C.I. Plosser. 1984. Money, credit, and prices in a real business cycle. *American Economic Review* 24: 363–380.

King, R.G. and S. Rebelo. 1986. *Business cycles with endogenous growth*. Unpublished working paper, University of Rochester.

King, R.G., and B. Trehan. 1984. Money: Endogeneity and neutrality. *Journal of Monetary Economics* 14: 385–394.

King, R.G., Plosser, C.I., Stock, J. and M. Watson. 1986. *Short run and long run relationships in macroeconomic time series*. Unpublished working paper, University of Rochester.

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

Leiderman, L. 1980. Macroeconomic testing of the rational expectations and structural neutrality hypotheses for the United States. *Journal of Monetary Economics* 6: 69–82.

Leiderman, L. 1983. The response of real wages to unanticipated money growth. *Journal of Monetary Economics* 1: 73–88.

Long, J.B., and C.I. Plosser. 1983. Real business cycles. *Journal of Political Economy* 91: 39–69.

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

Lucas Jr., R.E. 1972b. Econometric testing of the natural rate hypothesis. In *The econometrics of price determination*, ed. O. Eckstein. Washington, DC: Board of Governors of the Federal Reserve System.

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

Lucas Jr., R.E. 1977. Understanding business cycles. In *Stabilization of the domestic and international economy*, Carnegie-Rochester Series on Public Policy, vol. 5, ed. K. Brunner and A. Meltzer, 7–29. Amsterdam: North Holland.

Lucas Jr., R.E. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 2: 696–715.

Lucas, R.E., Jr. 1985a. *Models of business fluctuations*. Unpublished working paper, University of Chicago.

Lucas, R.E., Jr. 1985b. *On the mechanics of economic development*. Working paper, University of Chicago.

Mankiw, N.G. 1985. Small menu costs and large business cycles: A macroeconomic model of monopoly. *Quarterly Journal of Economics* 2: 529–538.

McCallum, B.T. 1978. Price level adjustments and the rational expectations approach to macroeconomic stabilization policy. *Journal of Money, Credit and Banking* 10: 418–436.

McCallum, B.T. 1979. A monetary policy ineffectiveness result in a model with a predetermined price level. *Economics Letters* 3: 1–4.

McCallum, B.T. 1980. Rational expectations and macroeconomic stabilization policy: An overview. *Journal of Money, Credit and Banking* 12(pt. 2): 716–746.

McCallum, B.T. 1986a. *On real and sticky price models of the business cycle*. National Bureau of Economic Research working paper No. 1933.

McCallum, B.T. 1986b. *On real and sticky price of the business cycle*. National Bureau of Economic Research working paper No. 1933.

Merrick Jr., J.J. 1983. Financial market efficiency, the decomposition of anticipated versus unanticipated money growth, and further tests of the relation between money and real output. *Journal of Money, Credit and Banking* 40: 222–232.

Mishkin, F.S. 1982. Does anticipated money matter: An econometric investigation. *Journal of Political Economy* 90: 22–51.

Mitchell, W.C. 1951. *What happens during business cycles*. New York: National Bureau of Economic Research.

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

Nelson, C.R., and C.I. Plosser. 1982. Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* 10: 139–162.

Phelps, E.S. 1967. Phillips curves, expectations of inflation and optimal unemployment over time. *Economica* 34: 254–281.

Phelps, E.S., and J.B. Taylor. 1977. Stabilizing powers of monetary policy under rational expectations. *Journal of Political Economy* 85: 163–190.

Phelps, E.S., and S.G. Winter. 1970. Optimal price policy under atomistic competition. In *Microeconomic foundations of employment and inflation theory*, ed. E.S. Phelps. New York: W.W. Norton and Company.

Prescott, E.C. 1986. Theory ahead of business cycle measurement. Federal Reserve Bank of Minneapolis Research Department Staff Report 102, February; also in *Carnegie-Rochester Conference Series on Public Policy.*

Rogerson, R. 1985. *Indivisible labor, lotteries and equilibrium*. Unpublished working paper, University of Rochester.

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

Rotemberg, J. 1982. Sticky prices in the United States. *Journal of Political Economy* 90: 1187–1211.

Sargent, T.J. 1973. Rational expectations, the real rate of interest, and the natural rate of unemployment. *Brookings Papers on Economic Activity* 2: 429–472.

Sargent, T.J. 1976. A classical macroeconometric model for the United States. *Journal of Political Economy* 84: 207–237.

Sargent, T.J. 1979. Granger causality and the natural rate hypothesis. *Journal of Political Economy* 87: 213–248.

Sargent, T.J. 1981. Interpreting economic time series. *Journal of Political Economy* 89: 403–410.

Sargent, T.J., and N. Wallace. 1975. Rational expectations, the optimal monetary instrument and the optimal

money supply rule. *Journal of Political Economy* 83(2): 241–254.
Solow, R.M. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39: 312–320.
Stigler, G., and J. Kindahl. 1970. *The behavior of industrial prices*. New York: Columbia University Press for the National Bureau of Economic Research.
Tinbergen, J. 1932. *Business cycles in the United States of America, 1919–1932: Statistical testing of business cycle theories*. Geneva: League of Nations.

Business Networks

Francis Bloch

Abstract

Informal and formal business networks play an increasing role in economic activities. Business networks have been studied both by sociologists and economists in order to answer three questions: What is the influence of business networks on economic activities? What are the determinants of business networks? When and how are business networks alternatives to organized markets?

Keywords

Business networks; Class; Commitment; Corporate governance; Credit markets in developing countries; Diffusion of innovations; Economic geography; Enforceable international contracts; Interlocking directorates; Intermediate goods; International trade; Knowledge spillovers; Labour market search; Market sharing; Matching; Spatial concentration; Strategic alliances

JEL Classifications

D85

Informal and formal business networks play an increasing role in economic activities. A large literature in economics and sociology has focused attention on these business networks. Three sets of questions have been raised: What is the influence of business networks on economic activities? What are the determinants of business networks? When and how are business networks alternatives to organized markets?

The importance of social networks has been stressed in three spheres of economic activities: the job market, where personal referrals play an essential role; international trade, where the existence of networks helps explain the volume of trade across borders; and urban economics, where business relations are an important determinant of the degree of local knowledge spillovers.

Empirical studies show that as many as half of the jobs are found through personal contacts. Granovetter's landmark study (1974) of the importance of networks in the managerial and professional job market in a Boston suburb stresses the difference between 'strong' and 'weak' ties. According to his study, weak ties – distant acquaintances to individuals who belong to different communities – play a much stronger role than 'strong ties' – close relations to individuals belonging to the same group – in helping business executives finding or changing jobs. Recent economic models of job networks emphasize the dynamic effect of networks on unemployment and inequality (Calvo-Armengol and Jackson 2004).

In international trade, informal co-ethnic networks formed by migrants – like the Chinese trading network – and formal business networks like the Japanese *keiretsu* have a significant impact on the volume of trade across borders. Rauch's survey (2001) summarizes the empirical evidence and outlines different theoretical explanations of the effect of business networks on international trade. The existence of personal links allows traders to match opportunities better as the network provides an informational link across agents from different countries. Networks also allow traders to solve the problems of enforcement of international contracts – agents who do not meet their obligations may be expelled from the network.

Informal networks also play a fundamental role in the diffusion of innovations and the emergence

of new ideas in local areas. In her celebrated comparison of business models in the Silicon Valley and on Route 128, Saxenian (1994) argues that the success of the Silicon Valley is in large part due to the flexible, informal organization of business relations in California. Economic geographers have long noted that these informal networks generate important knowledge spillovers, which help explain the concentration of industrial activities over space and justify the emergence of industrial districts.

The architecture of business networks has been extensively studied in two areas where precise data can be obtained: interlocking directorates and strategic alliances. Empirical studies of interlocking directorates – the exchange of directors across company boards – first show that networks of intercorporate relations are highly asymmetric: a small number of firms occupy a central position on the network, concentrating a large number of interlocks. Second, intercorporate links tend to be local, and interlocking occurs among firms in the same geographical area. Third, the number of interlocks increases with the firm's size.

To explain this pattern of interrelations, two competing theories have been proposed, resulting in a lively controversy in the sociological literature, reviewed by Mizruchi (1996). Proponents of the social class theory argue that interlocking reflects the dominance of the upper class, and that relations among firms are mostly explained by individual friendships and the desire to maintain hegemony over the corporate world. The resource dependence theory explains the existence of interlocks by the firms' desire to access resources detained by other firms. According to this theory, industrial companies exchange directors with financial institutions in order to obtain easier access to credit and with their suppliers in order to guarantee access to intermediate goods needed in production.

Strategic alliances are bilateral agreements among firms in the same industry. Agreements to launch joint R&D projects have received special attention in the literature. On the empirical side, a large database of bilateral research agreements has been developed by the MERIT center in Maastricht (Hagedoorn 2002). These data show a large increase in the number of partnerships in the 1990s, and demonstrate that firms increasingly use flexible contractual arrangements rather than joint-equity subsidiaries to launch new research programmes. Research partnerships are very unevenly distributed across industrial sectors, with high-tech industries (in particular information technology and the pharmaceutical industry) accounting for a very large share of agreements.

Goyal and Joshi (2003) propose a theoretical model to explain the formation of these collaborative networks. Their analysis explains the high density of the networks by showing that, in the absence of linking costs, firms always have an incentive to form strategic alliances. In the presence of linking costs, stable networks become asymmetric, with a small number of isolated firms facing a large group of interrelated firms. When firms choose their research investments after the network is formed, inefficiencies arise as firms have a tendency to fragment their investments over too many links. Belleflamme and Bloch (2004) study a different type of strategic alliance: reciprocal market-sharing agreements whereby firms divide markets geographically. They show that stable networks are typically asymmetric and contain complete components of different sizes.

Trade networks can provide a viable alternative to organized, anonymous, markets. Buyers and sellers establish personal links, and conduct trade on a bilateral basis rather than through a centralized market. Historically, business networks have played a fundamental role in the development of trade. Greif's celebrated study (1993) of the Maghribi network, formed by Jews in the western Mediterranean in early medieval Europe, points out that business networks were able to solve commitment problems in the absence of institutions enforcing contracts. Still in the western Mediterranean, but in modern times, Kirman's detailed study (2001) of the fish market in Marseille also shows that a larger volume of trade is conducted on a bilateral basis, with buyers and sellers linked through durable relations.

Casella and Rauch (2002) and Kranton (1996) propose alternative theoretical models to investigate the difference between anonymous markets and personalized networks. In Casella and Rauch (2002), business networks enable traders to overcome informational trade barriers, and to learn about matching opportunities in international markets. They show that agents who continue to conduct trade through organized markets suffer from the presence of the business network. Kranton's model (1996) is built around the issue of enforcement of contracts: agents can either choose to trade on the market at the risk of being cheated but benefiting from a wide variety of goods, or to use a personal network. Kranton shows that there exists a strong interaction between the two modes of exchange: the more people use networks, the lower their incentives to use markets; the larger the fraction of the population which uses markets, the lower are their incentives to engage in personal transactions.

In summary, the importance of business networks in economic activities, which has long been recognized by sociologists, is attracting increasing attention from economists. New theoretical and empirical methods enable researchers to revisit business networks. In this relatively new field of study, a number of problems remain open. For example, the theoretical corporate governance literature is still silent on the issue of interlocking directorates. The interaction between formal insurance and credit markets and informal network arrangements in developing countries also awaits further study.

See Also

- ▶ Corporate Governance
- ▶ Network Formation
- ▶ Social Networks in Labour Markets

Bibliography

Belleflamme, P., and F. Bloch. 2004. Market sharing agreements and stable collusive networks. *International Economic Review* 45: 387–411.

Calvo-Armengol, A., and M.O. Jackson. 2004. The effects of social networks on employment and inequality. *American Economic Review* 94: 426–454.

Casella, A., and J. Rauch. 2002. Anonymous markets and group ties in international trade. *Journal of International Economics* 58: 19–47.

Goyal, S., and S. Joshi. 2003. Networks of collaboration in oligopoly. *Games and Economic Behavior* 43: 57–85.

Granovetter, M.S. 1974. *Getting a job*. Cambridge, MA: Harvard University Press.

Greif, A. 1993. Contract enforceability and economic institutions in early trade: The Maghribi trader's coalition. *American Economic Review* 83: 525–548.

Hagedoorn, J. 2002. Inter-firm R&D partnerships: An overview of major trends and patterns since 1960. *Research Policy* 4: 477–492.

Kirman, A. 2001. Market organization and individual behavior: Evidence from fish markets. In *Networks and markets*, ed. A. Casella and J. Rauch. New York: Russell Sage.

Kranton, R.E. 1996. Reciprocal exchange: A self-sustaining system. *American Economic Review* 86: 830–851.

Mizruchi, M.S. 1996. What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. *Annual Review of Sociology* 22: 271–298.

Rauch, J. 2001. Business and social networks in international trade. *Journal of Economic Literature* 39: 1177–1203.

Saxenian, A. 1994. *Regional networks: Industrial adaptation in silicon valley and route 128*. Cambridge, MA: Harvard University Press.

Business Politics in the Gulf

Steffen Hertog

Abstract

This article discusses the political economy of state–business relations in the oil monarchies of the Arabian Peninsula. It explains the unusual position of the Gulf business class: GCC merchants are capital-rich and well established, yet they have become structurally isolated in the oil age, as their levels of state dependence are high and they have lost organic linkages to local citizens, since they pay almost no taxes and employ mostly foreigners. Historical and comparative dimensions of this setup are highlighted.

Keywords

Arabian Peninsula; Fiscal sociology; Gulf business; Political economy of the Middle East; Rentier state; State–business relations

JEL Classifications

D72; D78; N45; N55; N85; P48; Q32; Z13

This article discusses the political economy of state–business relations in the six oil-rich monarchies of the Gulf Cooperation Council (GCC): Bahrain, Kuwait, Oman, Qatar, Saudi Arabia and the United Arab Emirates (UAE). Business politics in the Gulf are unusual in several regards: on the one hand, GCC regimes are among the most pro-business in the developing world; local merchants have a long history of collective action; and local business possesses more managerial capacity and economic scale than its peers in most neighbouring Arab countries. Yet in today's GCC, business remains strongly state-dependent, both politically and economically, and is relatively isolated in the public sphere.

The following sections will provide a brief history of state–business relations in the Gulf, give an overview of the structural position of business relative to the state, and discuss the formal and informal institutions through which state–business relations are conducted. It ends with comparative remarks on what is typical and unusual about business politics in the Gulf.

History of State-Business Relations in the Gulf

There is significant continuity in the history of the GCC's capitalist classes: The private sector continues to be dominated by family businesses, which in many cases have existed for several generations. GCC rulers have historically taken a pro-trade position and have avoided anti-business policies like the nationalisations that happened in neighbouring Arab republics. Many social institutions, such as the informal 'majlis' meetings in which business people exchange information and discuss politics, have existed before the onset of oil. Some of the GCC ruling families themselves have a merchant background.

Yet the capitalist classes – called 'merchants' (*tujjar*) in local parlance due to their historical origins as traders – have witnessed a dramatic change below the surface during the 20th century. The roles of state and business have essentially been reversed: in the pre-oil age of the 19th century, the embryonic states of the Arabian Peninsula were fiscally dependent on merchant families, who enjoyed a relatively strong and autonomous social and economic position. Merchants were often the first to provide rudimentary utility, health and education services to local populations and financed rulers through taxes and loans. They enjoyed high geographic mobility, giving them a credible exit threat *vis-à-vis* rulers who did not pursue pro-merchant policies. Hence their political voice was powerful (Crystal 1995).

In the late 19th and early 20th centuries, the political autonomy of merchants and other social elites began to be circumscribed by the British colonial power, which buttressed and stabilised local rulers, giving them a measure of autonomy from local society (Onley and Khalaf 2006). The state, however, remained personalised, fiscally stretched and severely underdeveloped. The financial situation of local states would remain brittle, to the extent that, in the case of Qatar, the ruler at one point had to mortgage his house to avoid the collapse of his protogovernment. A more dramatic shift set in when rulers started receiving large-scale oil rents from foreign concessions after the Second World War. This allowed them to build large states, reduced their dependence on social elites and brought about the ascendancy of the state over all major social actors (Lienhardt 2001).

Merchant elites became clients of the state and the ruling families atop of it. As most large-scale business was now conducted with the state, they depended on government contracts, loans and subsidies. As non-oil taxes were abolished or

greatly receded in importance, business elites lost fiscal leverage over the state. Their exit threat lost credibility, as they would severely harm themselves through leaving – which in any case would not have made much of a difference to national economic development. As their own capacities to develop local infrastructure and supply modern goods were limited, they were in many cases reduced to the role of brokers between foreign contractors and suppliers on the one hand and the state on the other (Hertog 2010a, b). The merchant class lost much (albeit not all) of its capacity for collective action.

Individual merchants falling out of favour with rulers were quickly marginalised, while many new players from among the coterie of local ruling families entered the ranks of the merchant class (Field 1986). At the same time, members of the GCC's growing ruling families entered business themselves. The most prominent current example arguably is Mohammad bin Rashed of Dubai, but the Al Nahyan of Abu Dhabi, the Al Thani of Qatar, Saudi Arabia's Al Saud and, more recently Kuwait's Al Sabah have also staked out important positions in commerce. This has led to some crowding out of commoner merchants, especially in the lean years of the 1980s and 1990s, and particularly in the fields of commerce and real estate.

That being said, social links between rulers and merchants, albeit lopsided, have remained strong, and they remain prominent members of the GCC's state-dependent social elites. Merchant elites are consulted on economic policy matters and in some cases drawn on for governing, with representatives of leading merchant families typically holding at least a ministerial portfolio or two (Gause 1994; Peterson 2012). Private business has received strong state support in the shape of subsidies, targeted protection against foreign competition, and government loans. In all of these transactions, the state and its top elites are the more powerful actors, however; interactions between state and business tend to be more individualised than collective nowadays. Up until the present day, rulers have been able to 'make' new families out of nowhere by using discretionary patronage (see Azoulay 2012) on the phenomenon of state-sponsored Shiite tycoons in Kuwait).

B

Quantitative Indicators of Gulf Businesses' Current Structural Role

Macro-economic indicators underline the central role of the state and the relative marginalisation of private business in today's GCC. The private economy typically constitutes 30–40% of local GDPs and from the 1980s until the renewed oil-driven state spending boom of the late 2000s private business contributed the majority of national capital formation.

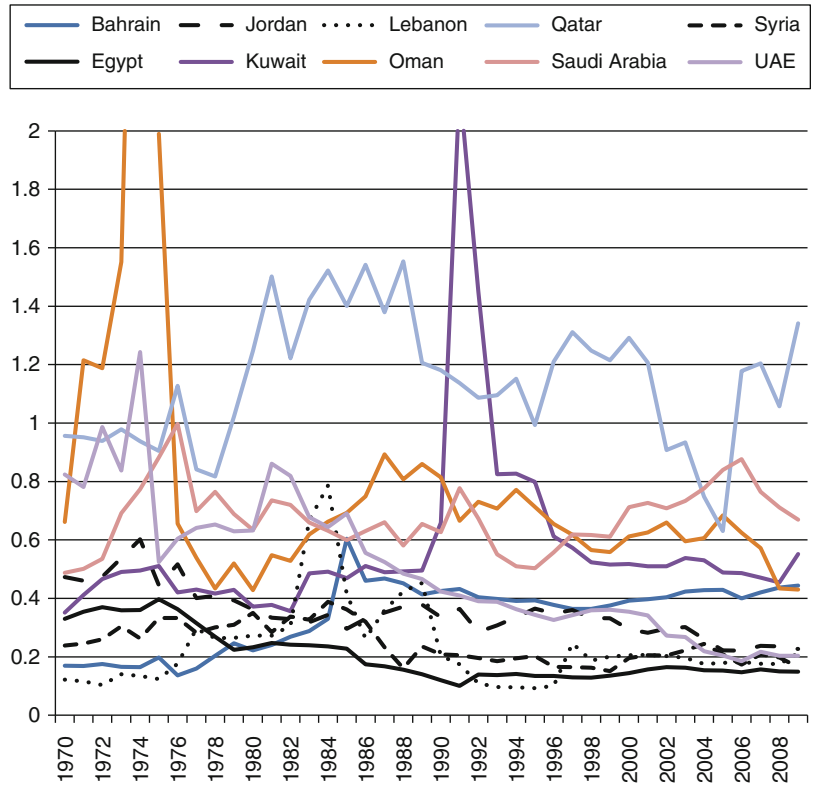
While this denotes substantial private activity, much of it consists of direct or indirect recycling of rents originating from the state (Hertog 2013). Even compared to the GCC's rather 'statist' Arab neighbours, GCC governments drive an unusually large share of consumption in the national economy, as Fig. 1 shows.

In addition, much of private household consumption is driven by government salaries and transfers to national households; privately earned salaries mostly accrue to foreign residents, who remit a large share of their earnings abroad.

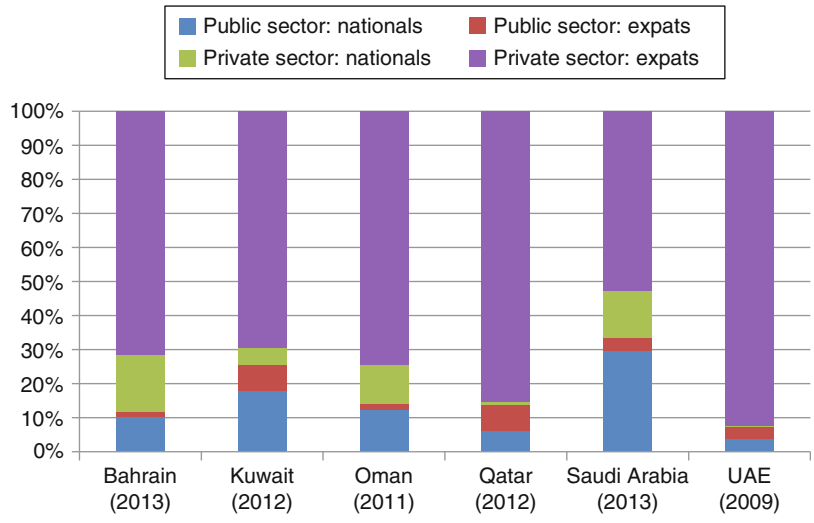
GCC business also contributes very little to the fiscal basis of local states (Beblawi and Luciani 1987). Most GCC countries levy no profit taxes on local companies, and none of them has a general sales tax or VAT. While most employment in the GCC is private, private sector jobs are predominantly held by foreigners, typically at much lower wages than what nationals earn in the public sector (Fig. 2).

The lack of national employment in the private sector means that a critical link between the citizenry and the local capitalist class is severed, reducing the shared economic interests between citizens and capitalists (Herb 2009).

Much of economic diversification in the GCC has moreover been driven by state-owned companies like SABIC, Emaar or Etisalat, rather than by private investors. The state-owned enterprises that



Business Politics in the Gulf, Fig. 1 Ratio of government to private consumption in GCC and regional neighbours (Source: UNSTATS) Note: The ratio for the UAE is likely underestimated, possibly due to lack of Emirate-level public consumption data



Business Politics in the Gulf, Fig. 2 Distribution of employment by sector and nationality in the GCC (Source: National agencies)

have been created since the first oil boom of the 1970s constitute the largest businesses and most coveted employers in the region and also tend to lead regional research and development efforts. Although managerially experienced and increasingly capital-rich (Luciani 2006), the technological capacity of private Gulf businesses is typically limited.

Private businesses in manufacturing also typically depend on the government's provision of cheap gas, electricity and fuel, as well as government-provided infrastructure and loans (Hertog 2013). It was also state-owned rather than private enterprises that have led the GCC push to export local services and goods on global markets, be they in logistics, telecoms, heavy industry, real estate or tourism.

Fiscal Sociology – GCC Merchants' Structural Isolation

The above structural background helps to understand the relative isolation of GCC business from other social and economic forces and the diminished political rule resulting from it, especially in the monarchies that enjoy a more open political space, notably Kuwait.

As indicated, processes that typically link the interests of citizens and business in non-rent economies have been severed or weakened in the GCC: Gulf business provides little employment for nationals, almost no taxes to finance public services and – as most wealth is held privately – also few investment opportunities for citizens (Herb 2009, 2014; Hertog 2013). This means that there is a potential zero-sum conflict between citizens and business over the use of state rents: Citizens will prefer subsidies and patronage for themselves rather than developmental spending that might benefit a business class whose growth yields few tangible benefits for the citizenry at large. General social modernisation has moreover increased GCC citizens' political awareness and reduced their respect for traditional social elites, who have lost some of their functions in the face of an omnipresent state.

These dynamics help to explain the increasing political marginalisation of business in the public political sphere, as well as an emerging anti-business populism that appears at odds with the strong pro-capitalist traditions of the region. The local press carries regular anti-business op-eds, blaming merchants for not employing locals, gazumping prices and not contributing enough to national development and social causes. GCC businessmen running for public office are regularly defeated in those cases where open elections are held; none of the GCC parliaments has a powerful pro-business bloc. Even in the tame appointed consultative councils of the UAE and Qatar anti-business sentiment has been visible at times.

Social modernisation and structural isolation of merchants are particularly pronounced in Kuwait, where citizens' political mobilisation is strong and the middle class well organised. Even more than their peers in the rest of the GCC, Kuwaiti merchants have been marginalised in public political life, notably in the powerful and openly elected parliament (Herb 2014). In the politically more quiescent UAE, by contrast, merchants have retained more of their social and political status (Almezaini 2012).

The decline in merchants' public political position is particularly striking given their critical role in first pushing local rulers towards creating elected assemblies in the 1920s, 1930s and 1950s and their prominent representation among the liberal-nationalist opposition of this earlier era in Bahrain, Kuwait and Dubai (Crystal 1995; Davidson 2007; Moore 2009).

As a result of their less prominent political position, the merchants' role in politics today is often reactive and defensive, focused on preserving their privileges and asking for state support. While generally in favour of bureaucratic reform and regulatory streamlining, merchants have tended to resist the opening of new sectors to foreign investors, which has been a key component of GCC efforts to turn local markets into regional and global economic hubs.

Although running larger-scale operations and possessing deeper capital resources than their peers in most of the poorer Middle East and North Africa (MENA) states, Gulf business classes are less politically active and visible. In recent years, governments have put them under strong pressure to step up their contribution to the employment of nationals, notably in the case of Saudi Arabia's 'Nitaqat' policy, which has imposed 'Saudization' quotas on private employers since 2011. Despite considerable costs, Saudi business has largely fallen in line with this policy, given the strong public criticisms and royal pressure to which it has been subjected.

The Organisation of State–Business Relations

The reactive nature of business lobbying is to some extent also explained by the internal structure of the GCC business class: it is stratified into a layer of elite national families on the one hand and a large number of small and medium enterprises, on the other hand, which are both nationally and foreign-owned. Smaller businesses are often unrepresented in chambers of commerce, which tend to be dominated by a few large families, further undermining the corporate coherence of business.

Elections for the boards of chambers usually witness low turnout. While the government consults chambers on economic policy drafts, these institutions often lack technical capacity for specialised policy-making. Below the peak level of chambers of commerce and industry, the GCC has few powerful sector-specific associations, again lagging behind the organisation of business in many neighbouring Arab countries. While prominent merchants often enjoy informal access to ruling elites, this can be used to pursue individual agendas as much as for the representation of the merchant class at large. It is hence not always clear whether such access helps or hinders collective bargaining between state and business.

Elite business families sometimes enjoy representation on the top level of government. This, however, is the case less often than used to be the case in the early era post-Second World War of state-building: ruling elites have a wider pool of non-merchant technocrats to choose from than in the days when the scions of merchant families often were the only nationals who could afford foreign university education. Merchants remain relatively better represented in the UAE and Qatari governments, where politics is more elite-dominated and popular pressures are more muted (Almezaini 2012). A similar situation existed in Oman until 2011, when public protests in the wake of the Tunisian and Egyptian uprisings led Sultan Qaboos to fire several long-term representatives of the merchant class from his cabinet (Valeri 2012).

Comparative Remarks

To which extent do state–business relations in the GCC fall in line with broader patterns of business politics in the developing world? State dependence, clientelism and prominence of kinship networks are not unusual in less developed and emerging markets (Heydemann 2004; Maxfield and Schneider 1997), even if the level of structural dependence on the state in the GCC appears particularly high. What is more particular about the GCC is the contrast between the rather high managerial and production capacity of business on the one hand and its isolated political position on the other.

In comparison with other developing countries, Gulf business classes look back on a fairly continuous trajectory of development, owing to the consistently pro-capitalist and conservative orientation of GCC ruling elites since before the oil age. The private sector continues to play a prominent role in national developmental planning and rhetoric. Yet they are *de facto* marginalised in the public political sphere as they lack organic links to much of the citizenry, a pattern that is largely explained with the rentier nature of Gulf regimes in which citizens depend on the state much more than on business. Only in more authoritarian GCC countries like the UAE is the government able to pursue consistently pro-capitalist policies (Herb 2014). Future research might show whether these patterns also apply in other high-rent countries outside of the GCC region.

Outlook

The political isolation of GCC business has been a secular trend of the post-Second World War era. It is not clear whether this trend will continue. On the one hand, citizens' political awareness is on the rise across the region. New generations are less likely to defer reflexively to established social elites if these cannot justify their position through contributions to national development that provide tangible benefits to citizens. On the other hand, the increased pressure for the private sector to employ nationals could lead to higher citizen participation in the

private labour market and hence create a growing popular constituency in favour of sound economic policies that benefit business. It is in the long-term collective self-interest of GCC business to shift towards national employment. It does, however, create short-term costs and is subject to severe collective action problems.

The overshadowing of the private sector by the state in the GCC since the early 2000s is also rooted in rising oil revenues that have allowed governments to set the pace of economic development. What if oil prices stagnate or fall again? When this happened in the 1980s and 1990s business suffered, yet became relatively more important to cash-strapped governments as it was needed for financing non-oil diversification. Many of the liberalising economic reforms of the 2000s had their roots in the austerity of the 1990s. Unless austerity should lead to wider political instability, such a dynamic could unfold again. Although the merchant class is unlikely ever to regain the political position of the pre-oil age, greater employment of nationals and future austerity could bring it somewhat closer to its pre-oil status.

See Also

- ▶ Algeria, Economy of
- ▶ Egypt, Economy of
- ▶ Jordan, Economy of
- ▶ Libya, Economics of
- ▶ Oil and Politics in the Gulf: Kuwait and Qatar
- ▶ Oman, Economy of
- ▶ Syria, Economy of
- ▶ Tunisia, Economy of
- ▶ Yemen, Economy of

Bibliography

Almezaini, K. 2012. Private sector actors in the UAE and their role in the process of economic and political reform. In *Business politics in the Middle East*, ed. S. Hertog, G. Luciani, and M. Valeri. London: C Hurst & Co.

Azoulay, R. 2012. The politics of Shi'i merchants in Kuwait. In *Business politics in the Middle East*, ed. S. Hertog, G. Luciani, and M. Valeri. London: C Hurst & Co.

Beblawi, H., and G. Luciani (eds.). 1987. *The Rentier state*. London/New York: Croom Helm.

Crystal, J. 1995. *Oil and politics in the Gulf: Rulers and merchants in Kuwait and Qatar*. Cambridge: Cambridge University Press.

Davidson, C.M. 2007. Arab nationalism and British opposition in Dubai, 1920–66. *Middle Eastern Studies* 43(6): 879–892. doi:10.1080/00263200701568246.

Field, M. 1986. *The merchants*. New York: Overlook Press.

Gause, F.G. 1994. *Oil monarchies*. New York: Council on Foreign Relations Press.

Herb, M. 2009. A nation of bureaucrats: Political participation and economic diversification in Kuwait and the United Arab Emirates. *International Journal of Middle East Studies* 41(3): 375–395. doi:10.1017/S0020743809091119.

Herb, M. 2014. *The wages of oil: Parliaments and economic development in Kuwait and the UAE*. Ithaca: Cornell University Press.

Hertog, S. 2010a. The sociology of the Gulf rentier systems: Societies of intermediaries. *Comparative Studies in Society and History* 52(2): 282–318. doi:10.1017/S0010417510000058.

Hertog, S. 2010b. *Princes, brokers, and bureaucrats: Oil and the state in Saudi Arabia*. Ithaca: Cornell University Press.

Hertog, S. 2013. *The private sector and reform in the Gulf Cooperation Council*. Working paper. Kuwait Program, London School of Economics, London.

Heydemann, S. (ed.). 2004. *Networks of privilege in the Middle East: The politics of economic reform revisited*. Basingstoke: Palgrave Macmillan.

Lienhardt, P. 2001. *Shaikhdoms of Eastern Arabia*. Basingstoke: Palgrave, in association with St. Antony's College, Oxford.

Luciani, G. 2006. From private sector to national bourgeoisie: Saudi Arabian Business. In *Saudi Arabia in the balance: Political economy, society, foreign affairs*, ed. P. Aarts and G. Nonneman, 144–181. New York: NYU Press.

Maxfield, S., and B.R. Schneider (eds.). 1997. *Business and the state in developing countries*. Ithaca: Cornell University Press.

Moore, P.W. 2009. *Doing Business in the Middle East: Politics and economic crisis in Jordan and Kuwait*. Cambridge: Cambridge University Press.

Onley, J., and S. Khalaf. 2006. Shaikhly authority in the pre-oil Gulf: An historical–anthropological study. *History and Anthropology* 17(3): 189–208. doi:10.1080/02757200600813965.

Peterson, J. 2012. Rulers, merchants and Shaikhs in Gulf politics. In *The Gulf family: Kinship policies and modernity*, ed. A. Alsharekh, 21–36. London: Saqi.

Valeri, M. 2012. Oligarchy vs. oligarchy: Business and the politics of reform in Bahrain and Oman. In *Business politics in the Middle East*, ed. S. Hertog, G. Luciani, and M. Valeri. London: C Hurst & Co.

Butlin, Noel George (1921–1991)

Graeme Donald Snooks

Keywords

Butlin, N. G.; Economic development; National accounting; Neoclassical growth theory; Structural disequilibrium; Technical change

JEL Classifications

B31

Noel George Butlin, one of Australia's leading historical economists, was born in Singleton, New South Wales on 19 December 1921. He was the sixth child and third son of Thomas Lyon Butlin, a railway porter, and Sara Mary Butlin (née Chantler). Butlin attended Maitland Boys High and studied economics at Sydney University. During his undergraduate years, Sydney had the nation's best economics department in terms of the professional qualifications of its teaching staff. Even so, Butlin claimed that, while his lecturers taught him how to deconstruct aspects of the economy, they were unable to show him how it all worked. He wanted to become a scholar to understand real-world economic processes.

Like many others of his generation, Butlin's career was disrupted by war. While he wanted to enter academia, the only avenue available on graduation was the Australian public service. Between 1942 and 1945 Butlin was mainly seconded to posts in the UK and USA. There he met J.M. Keynes, L. Robbins, A. Robertson, R. Stone, and H.J. Habakkuk. Back in Australia he participated in 1945 in making plans for Australia's post-war reconstruction, and in 1946 finally took up a lectureship at Sydney University. To further his research ambitions, Butlin accepted a Rockefeller Fellowship in 1949 to study for a Ph.D. at Harvard under Joseph Schumpeter. Unfortunately, the great man died a few months after Butlin's arrival, and he found himself in Harvard's Centre for Entrepreneurial Studies. He had little sympathy with their growing sociological interests and, after initial research on Canadian railways, decided in 1951 to return to Australia to work at the Australian National University (ANU). In 1963 he became Professor and Head of the Department of Economic History. Butlin's 40-year association with the ANU ended only with his death on 2 April 1991.

Back in Australia, Butlin was swept up in the post-war concern with economic development. On the theoretical side, the old influence of Schumpeter was joined by the new influences of Harrod and Solow–Swan, and, on the measurement side, the great statistician Coghlan was joined by Kuznets. Butlin absorbed ideas from them all. He borrowed the 'structural disequilibrium' concept from Schumpeter, but ignored technological change in favour of the investment focus of the neoclassical growth model. Economic development in Butlin's analysis proceeded via long investment booms that created structural disequilibria and required depressions to reattain structural balance. The outcome of these influences, together with much hard work during the 1950s, was the publication of his two-volume magnum opus on Australian development (Butlin 1962, 1964). This set the pattern for subsequent analysis by historians and economists in the 1970s and 1980s, and was only challenged in the 1990s (Snooks 1994). Despite being an active researcher until his death, Butlin never surpassed this early work. His most interesting subsequent research focused on pushing his GDP estimates back to 1788 (Butlin 1986), and on analysing the Aboriginal economy (Butlin 1983, 1994).

What was the nature and importance of Butlin's contribution to economics and history? First and foremost, Butlin focused our attention on the *process* of Australian economic development, and showed that it was endogenously generated. This was an essential counterpoint to the traditional view that development was exogenously driven. Second, he demonstrated that

real-world growth processes could not be encompassed by the simple neoclassical growth models that were fashionable among orthodox economists at the time. Unfortunately he was unable to fulfil his intention of writing a 'strictly analytical' volume to complete the 1960s trilogy. He failed, therefore, to develop a general dynamic theory that could displace these totally unrealistic growth models. That was left to others (Snooks 1998). Third, while his hybrid national accounting techniques have been criticized, they have weathered the storm reasonably well. More than most Australian national accountants, Butlin had an impressive understanding of the history that generated the data he employed. When used for long-run rather than year-to-year analysis, the differences in alternative estimates are not significant (Snooks 2007). In any case, it is Butlin's overarching interpretation, his realist vision, and his important example of what can be done with the available data that constitute his enduring contribution.

See Also

► Australasia, Economics in

Selected Works

1962. *Australian Domestic Product, Investment and Foreign Borrowing, 1861–1938/ 39*. Cambridge: Cambridge University Press.

1964. *Investment in Australian Economic Development, 1861–1900*. Cambridge: Cambridge University Press.

1983. *Our Original Aggression: The Aboriginal Populations of Southeastern Australia, 1788–1850*. Sydney: Allen & Unwin.

1986. *Contours of the Australian economy, 1788–1860*. Australian Economic History Review 26, 96–125.

1994. *Economics and the Dreamtime*. Cambridge: Cambridge University Press.

Bibliography

Snooks, G.D. 1994. *Portrait of the family within the total economy: A study in longrun dynamics, Australia, 1788–1990*. Cambridge: Cambridge University Press.

Snooks, G.D. 1998. *Longrun dynamics: A general economic and political theory*. London: Macmillan.

Snooks, G.D. 2007. Dynamics downunder: Australian economic strategy and performance from the palaeolithic to the twenty-first century. In *World economic performance: Past, present and future*, ed. P. Rao and B. van Ark. Princeton: Princeton University Press.